

Statistical Approaches to Inferring Object Shape from Single Images

By

Ashwini Shikaripur Nadig

Submitted to the Department of Electrical Engineering and Computer Science and the
Graduate Faculty of the University of Kansas
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy

Dr. Bo Luo, Chairperson

Dr. Brian Potetz, Co-chair

Committee members

Dr. Luke Huan

Dr. James Miller

Dr. Paul Selden

Date defended:

May 20, 2014

The Dissertation Committee for Ashwini Shikaripur Nadig certifies
that this is the approved version of the following dissertation :

Statistical Approaches to Inferring Object Shape from Single Images

Dr. Bo Luo, Chairperson

Date approved: May 20, 2014

Abstract

Depth inference is a fundamental problem of computer vision with a broad range of potential applications. Monocular depth inference techniques, particularly shape from shading dates back to as early as the 40's when it was first used to study the shape of the lunar surface. Since then there has been ample research to develop depth inference algorithms using monocular cues. Most of these are based on physical models of image formation and rely on a number of simplifying assumptions that do not hold for real world and natural imagery. Very few make use of the rich statistical information contained in real world images and their 3D information. There have been a few notable exceptions though. The study of statistics of natural scenes has been concentrated on outdoor scenes which are cluttered. Statistics of scenes of single objects has been less studied, but is an essential part of daily human interaction with the environment. Inferring shape of single objects is a very important computer vision problem which has captured the interest of many researchers over the past few decades and has applications in object recognition, robotic grasping, fault detection and Content Based Image Retrieval (CBIR). This thesis focuses on studying the statistical properties of single objects and their range images which can benefit shape inference techniques. I acquired two databases: Single Object Range and HDR (SORH) and the Eton Myers Database of single objects, including laser-acquired depth, binocular stereo, photometric stereo and High Dynamic Range (HDR) photography. I took a data driven approach and studied the statistics of color and range images of real scenes of single objects along with whole 3D objects and uncovered

some interesting trends in the data. The fractal structure of natural images was previously well known, and thought to be a universal property. However, my research showed that the fractal structure of single objects and surfaces is governed by a wholly different set of rules. Classical computer vision problems of binocular and multi-view stereo, photometric stereo, shape from shading, structure from motion, and others, all rely on accurate and complete models of which 3D shapes and textures are plausible in nature, to avoid producing unlikely outputs. Bayesian approaches are common for these problems, and hopefully the findings on the statistics of the shape of single objects from this work and others will both inform new and more accurate Bayesian priors on shape, and also enable more efficient probabilistic inference procedures.

Acknowledgements

I would like to express a deep sense of gratitude and admiration to my advisor, Dr. Brian Potetz for his invaluable guidance and support throughout my graduate studies at the University of Kansas. I would like to specially thank him for continuing to mentor and guide me through my research despite his full time schedule after leaving the university. But for his excellent support, I wouldn't have come so far. He is an incredible mentor and a perfect role model to me, his patience and dedication towards his students and research is an inspiration that I'll carry forward through all my future endeavors. I would like to extend a special thanks to Dr. Bo Luo for his support in the process. I also thank Dr. James Miller, Dr. Luke Huan and Dr. Paul Selden for serving on my committee. I would like to thank my husband, Anvesh, for being so supportive and providing me with encouragement in times when I needed it most. I would like to thank my parents, my sister Pallavi and brother-in-law Karthik who always believed in me and supported me all along my academic career.

Contents

1	Introduction	1
1.1	Overview	3
2	Background	4
2.1	Scene Statistics	4
2.2	Shape from Shading	6
2.2.1	Mathematical model	6
2.2.2	Common Approaches	7
2.2.3	Linear Techniques:	8
2.2.4	Pentland's approach:	9
2.2.5	Tsai and Shah's method	10
2.3	Shape from Shadow	11
2.4	Statistical approaches to depth inference	12
3	Datasets of Single Objects with Ground-Truth Depth	15
3.1	Introduction	15
3.2	Single Object Range and HDR (SORH) Database	17
3.2.1	Acquisition Process of HDR Images and Range Scans	17
3.2.2	High Dynamic Range (HDR) imaging	19
3.2.3	Photometric Stereo	24
3.2.4	Binocular Stereo	26

3.3	Eton Myers Database	28
4	The statistics of the 3D shape and appearance of single objects and surfaces	30
4.1	Introduction	30
4.2	Background	32
4.3	A Scaling Hypothesis for the Range Images of Single Objects	38
4.3.1	A Derivation of the Consequences of Self-Similarity	38
4.3.2	A Theoretical Justification of $1/f^4$ Power Spectra of Range Within Occlusion Boundaries	41
4.4	Scaling properties of unoccluded surfaces	46
4.5	Study of scaling properties of 3D objects	48
4.6	Discussion	52
5	Statistical Correlations Between Intensity and Range Image Pairs of Single Objects	56
5.1	Introduction	56
5.2	Mathematical Background	57
5.2.1	Linear Regression Model	57
5.2.2	Pentland's Linear Lambertian Model	58
5.3	Experimental Methodology	60
5.3.1	The patch based approach	61
5.3.2	The pair-wise approach	62
5.4	Depth Inference Using Ridge Regression	66
6	SFS Using Expectation Propagation	69
6.1	Probabilistic Graphical Models in Vision	69
6.2	Overview of Markov Random Fields	69
6.3	Overview of Expectation Propagation	71

6.4	Whitened Expectation Propagation	74
6.5	Whitened EP for Shape From Shading (SFS)	76
6.6	Implementation details	77
6.6.1	My approach: Learning potentials	77
6.6.1.1	Similarity measures	79
6.7	Discussion and open questions	84
A	SORH Imagery	100

List of Figures

3.1	Example objects from the SORH database	17
3.2	Approximate positions of the camera (C), object (O) and the nine light sources.	19
3.3	Intial experimental setup for laser range scanning.	20
3.4	The retrieved response functions for the red, green and blue channels of the Nikon D90 DSLR camera.	22
3.5	Example objects from the Eton Myers database	28
4.1	Example collage model image with $1/r^3$ distribution of object radii and $1/f^2$ power spectrum.	34
4.2	Plot showing the average power spectrum along with the power law fits for the SORH and Eton Myers images	47
4.3	Plot showing average power spectra of Eton Myers objects vs. the power spectra of an artificial scale-invariant object	51
5.1	Color coded 2D rendering of the correlation between intensity at pixel (13, 13) and all the pixels of the range patch for the 9 light source directions for the SORH dataset	63
5.2	Color coded 2D rendering of the correlation between intensity and range pixels separated by distances ranging from 0 to 100 in either direction for the 9 light source directions for SORH dataset	64

5.3	Color coded 2D rendering of the correlation between intensity and range pixels separated by distances ranging from 0 to 100 in either direction for the Eton Myers dataset	65
5.4	Range images of ground truth (first column) and predicted shape (second column) using linear regression of two objects from the SORH database. . .	66
5.5	3D mesh plots of ground truth (first column) and predicted shape (second column) using linear regression of two objects from the SORH database. . .	67
6.1	a) Query intensity patch, b) and c) corresponding query range and label patches. The valid and invalid pixels as well the boundary separating them are clearly marked in the label patch. d) Top 25 intensity patch matches along with their e) corresponding range patches.	72
6.2	Top row: Sample intensity patches. Bottom row: Kernel density estimates for the pixel centered at the corresponding patch in the top row.	81
6.3	a) Color images of test SORH objects, b) ground truth range images, c) predicted range images using $1/f^3$ EP	83
6.4	3D mesh plots of predicted shape using EP using a $1/f^3$ covariance structure, for two objects from the SORH database in 3 rotated views.	85
6.5	3D mesh plots of predicted shape using EP using a $1/f^4$ covariance structure, for two objects from the SORH database in 3 rotated views.	86

List of Tables

6.1	Comparison of results for linear regression and whitened EP. Each row corresponds to the percentage of surface normals that lie within the given angular separation of the ground-truth surface normal.	80
-----	---	----

Chapter 1

Introduction

The human visual system is adept at inferring 3D shape of objects from flat 2D scenes. It makes use of binocular cues such as stereo and monocular cues such as shading, texture, specularities, motion, contour, occlusion, shadow, perspective, haze, etc. Understanding and inferring the 3D structure of a scene is a fundamental problem of computer vision and there have been many psychophysical studies to understand how humans perceive shape. Ramachandran [1] states through experiments that shading provides one of the strongest cues of 3D shape along with outline, surface reflectance and motion. There are two methods of shape inference from shading cues: Shape from Shading (SfS) which uses just one image and photometric stereo which uses multiple images taken under different lighting conditions. Woodham [2] first introduced photometric stereo. Classical SfS was formulated by Horn [3] in the late 70s. It has been over four decades and SfS still remains to be a challenging problem with lots of scope for research. A majority of SfS algorithms start with a physics based model of image formation. These models are then inverted to solve for depth. A major downside to this is the sheer number of simplifying assumptions made like Lambertian reflectance, distant point light source with known direction, uniform albedo and no cast shadows to make the problem tractable. Even after these assumptions SfS remains to be highly ambiguous and nonlinear, yielding poor results even for synthetic imagery. Moreover, these assumptions are

highly unrealistic as natural scenes are characterized by diffuse lighting, variations in albedo, complex cast shadows and interreflections making SfS algorithms brittle towards such scenes.

Most of the popular vision algorithms for depth inference do not tap into the rich statistical information contained in natural scenes and their 3D structure. Understanding the statistics of natural scenes and their shape can not only yield strong depth cues which can be incorporated into depth inference algorithms, but can also further our understanding of the human visual system. The major advantage of statistical approaches are that they are not limited to any particular image formation phenomenon or shading cues alone. They can even encode other cues like shadow, occlusion, texture, perspective, etc. Natural scenes contain numerous statistical regularities that cannot be captured by physical models of image formation. These regularities can be observed in the natural geometry and arrangement of objects (for example, outdoor scenes are usually characterized by a horizontal ground plane and other things like trees and houses sticking vertically with respect to the ground), the distributions of size and shape of objects in space, the relative position of the observer and the natural distributions of light which can be exploited by statistical approaches. It took the human visual system years of constant learning right from the time of birth to be able to perceive complex scenes and shapes. One of the most highly exploited regularities by our visual system is the tendency to assume that light comes from above [1]. Another regularity is that nearer regions appear brighter and farther regions and regions in crevices and concavities tend to be darker [4] which was also highly exploited by artists in the past to convey illusions of depth. Although there has been significant interest in the study of statistics of cluttered natural scenes, little is known about the statistics of appearance or the shape of single objects. One possible reason for this might be the lack of real world datasets of single objects while datasets of natural scenes are readily available. Yet, single objects play a dominant role in daily human interaction with the environment. The problem of inferring 3D shape of a single object has applications in object recognition, robotic grasping, fault inspection, Content Based Image Retrieval (CBIR), etc. Such algorithms can benefit from

strong probabilistic priors over appearance and shape which can be learned from real-world data. To aid in my study, I acquired two new databases of range scans and color images of a variety of natural and man-made single objects viewed from a fixed distance. This thesis focuses on using these databases to study the statistical trends of the 3D shape of real-world single objects directly. These trends are leveraged to develop a superior Expectation Propagation shape inference algorithm [5].

1.1 Overview

The rest of the thesis is organized in the following way:

- Chapter 2 deals with a brief overview of the background, a mathematical formulation of traditional shape from shading and motivation for statistical approaches to solve the problem.
- Chapter 3 describes the Single Object Range and HDR (SORH) and Eton Myers datasets in detail; including the motivation for acquiring the datasets, acquisition setup and process.
- Chapter 4 investigates into the statistics of the 3D shape of single objects from the datasets. Both analytical and experimental proofs show that range images of single objects without self occlusion and observed under orthographic projection show scale invariant properties. Whole 3D objects on the other hand can be modeled as simple 3D shapes with overlaid scale invariant texture.
- Chapter 5 investigates into the correlational statistics of range imagery and a simple linear ridge regression analysis conducted on the single object datasets.
- Chapter 6 applies the statistical insights about the 3D world gained in chapter 4 to improve an Expectation Propagation (EP) algorithm for shape inference. Results of shape inference using EP are discussed.

Chapter 2

Background

2.1 Scene Statistics

This section provides a broad overview of the research that has been conducted to explore scene statistics. For our purposes, we can categorize the study of scene statistics into three areas: natural image statistics, range image statistics and joint statistics between natural images and their corresponding depth information.

The usual approach to shape from shading has involved devising models based on the physics of image formation and then inverting these models to solve for depth. Such models are highly underconstrained requiring various assumptions such as lambertian reflectance, uniform albedo, single point light sources, lack of cast shadows, etc. Such assumptions restrict the applicability of these techniques to a very narrow subset of images for which these assumptions hold; making generalization to unrestricted natural imagery difficult. Very few techniques have attempted to relax these tight assumptions but have not been very successful. There is a need to go beyond this trend and gain some understanding of the statistical relationships that exist between 2D images and their corresponding range images which can be incorporated into current shape inference algorithms. There has been a lot of interest in natural image statistics and the perception of natural signals [6–11]. The

most significant among these is the property of scale invariance of natural images which is quantified by the $1/f^2$ distribution of the power spectrum [7,8]. Another observation is that the histograms involving linear filters are highly kurtotic indicating non-Gaussianity and can be modeled using a generalized Laplace distribution [12]. The distributions of contrasts and contrast gradients are also highly non-Gaussian [7]. Although these statistics are highly variable from single image to image, taken as an ensemble are highly robust. Another key observation is that the independent components of natural scenes are Gabor-like edge filters whose responses are highly kurtotic [9]. The statistics of range images are very similar to that of natural images (non-Gaussianity, scale invariance, etc) [13] and can be represented by the "collage-model" for natural images studied in [14]; a notion that the world can be broken down into discrete objects of varying sizes which occlude each other. Another study of range images was performed in [15] where statistics like scene roughness, size-distance relationships, surface orientation, local curvature, and independent components of natural scenes were studied. Despite the tremendous interest in natural image statistics, the study of joint statistical trends between natural scenes and their depth information and how humans perceive depth is still in its nascent stage [4,16,17]. A few of them are reviewed below.

Torralba and Oliva [17] state that there are structural regularities in both man-made and natural images which differ with scale and by studying such properties we can infer the scale and absolute depth of the scene. In [18], the authors describe an abstract of the scene which they call "spatial envelope" using perceptual dimensions like openness, roughness, etc which can be estimated. Torralba and Freeman [19,20] proposed a concept called "Shape Recipes" which are functions that describe the relationship between an image and its shape. Basically, they learn the relationship between the image and shape at low-resolution and apply it at high-resolution to improve stereo estimates making an assumption that the shape recipes vary slowly over scale. Potetz and Lee [21] extend the concept of shape recipes by stating that the shape recipe kernels can be actually fit by a $1/r$ power-law which is also explained by the linear Lambertian model of shading under oblique lighting conditions. In

[4], Potetz and Lee study the statistical correlations between image and depth and make some interesting observations; an absolute correlation between nearness and brightness which they call "da Vinci correlation". They also attribute the power-law behavior of the real part of the shape recipe kernel to this correlation which was otherwise contradictory to the linear Lambertian model. The cast shadows and lighting interreflections due to diffuse lighting conditions which are ignored due to mathematical difficulties are responsible for the "da Vinci correlation". They show that the shadow cues contribute significantly more than just the shading cues for both urban and rural scenes taken under diffuse lighting conditions.

Until now, studies of statistics of natural scenes have been focused on cluttered outdoor scenes with many objects occluding each other with lots of scope for shadows. Statistics in such scenes are highly regular with a ground plane and objects sticking out of it perpendicularly. Little is studied about statistics of single objects which play an important role in our daily interaction with the environment. One possible reason might be the lack of databases of real world single objects, whereas databases of cluttered natural scenes are in plenty. An understanding of statistics of single objects will benefit many vision applications like shape inference, content based image retrieval, image denoising, object recognition etc. Section 2.2.1 explains the mathematical model of shape from shading and sections 2.2.4, 2.2.5 go through linear approaches to shape from shading.

2.2 Shape from Shading

2.2.1 Mathematical model

The main goal of Shape from Shading is to reconstruct the 3D shape from a given single 2D intensity image. The problem was first formalized by Horn as a solution of a first-order nonlinear partial differential equation with two unknowns [3]. Mathematically, if a surface is defined as $z = z(x, y)$ and $p = \frac{\partial z(x, y)}{\partial x}$ and $q = \frac{\partial z(x, y)}{\partial y}$ are the gradients in the x and y directions respectively, the image intensity $I(x, y)$ and the reflectance map $R(p, q)$ are related

as follows:

$$I(x, y) = R(p, q) \quad (2.1)$$

Clearly, the above equation is underconstrained with more number of unknowns than equations. So there is a necessity for additional constraints without which there is no unique solution.

For a Lambertian surface illuminated by a distant point light source, the reflectance map is given by:

$$R(p, q) = \rho \vec{N} \cdot \vec{s} \quad (2.2)$$

where ρ is the albedo of the surface, \vec{N} is the unit normal to the surface given by:

$$\vec{N} = \frac{1}{\sqrt{1 + p^2 + q^2}}(-p, -q, 1)^T \quad (2.3)$$

and \vec{s} is the unit vector pointing towards the light source which is assumed to be known:

$$\vec{s} = \frac{1}{\sqrt{1 + p_s^2 + q_s^2}}(-p_s, -q_s, 1)^T \quad (2.4)$$

and hence the reflectance map is:

$$R(p, q) = \frac{\rho(1 + pp_s + qq_s)}{\sqrt{(1 + p^2 + q^2)}\sqrt{(1 + p_s^2 + q_s^2)}} \quad (2.5)$$

2.2.2 Common Approaches

A lot of techniques for shape from shading have been developed since the 70's but none have been successful in recovering shape from unrestricted natural imagery. All the seminal work is compiled by Brooks and Horn [22]. Numerous techniques have been developed to

solve SfS and readers are directed to [23, 24] for a more comprehensive overview on the subject.

Most of the techniques developed so far rely on a number of simplifying assumptions:

- It is assumed that all surfaces can be modeled using a Lambertian reflectance model. Hence transparent, rough or shiny surfaces are ruled out.
- The light source is assumed to be a distant point source without any interreflections or cast shadows from other objects in the scene.
- The surface is assumed to have uniform albedo/color without any markings or texture.
- Orthographic projection is assumed most of the times with an exception in [25].

Although the simplifying assumptions greatly reduce the ambiguity and mathematical complexity of the problem, it still remains highly ambiguous and nonlinear. Moreover the assumptions highly restrict the applicability of these algorithms to a limited set of images, mostly synthetic. There are some methods which seek to reduce the ambiguity by assuming that the true depth is known along the border of the image or at certain "singular points" in the image [24]. Another class of approaches seek to minimize one or more energy functions across the image either by calculating the Euler-Lagrange equations or by directly minimizing them [24, 26, 27]. The energy functions can be brightness constraints, integrability constraints, smoothness constraints, etc.

In the recent past, there have been many approaches that seek to estimate the rough scene geometry from outdoor images [28–31]. They assume that the camera is oriented in parallel to the ground and focus on estimating only the rough scene geometry and not the fine detail from close-ups of any objects in the scene.

2.2.3 Linear Techniques:

There are a plethora of techniques to solve the SFS problem but none generalize to natural imagery so far. This section discusses the two popular linear approaches to solve the

problem: Pentland's approach [32] and Tsai-Shah's approach [33]. The following sections will go through each in detail.

2.2.4 Pentland's approach:

Pentland [32] introduced a technique which forms a linear approximation of the reflectance function in terms of p and q . This is done by taking a Taylor expansion of R around $(p, q) = (p_0, q_0)$:

$$R(p, q) = R(p_0, q_0) + (p - p_0) \left. \frac{\partial R(p, q)}{\partial p} \right|_{p=p_0, q=q_0} + (q - q_0) \left. \frac{\partial R(p, q)}{\partial q} \right|_{p=p_0, q=q_0} \quad (2.6)$$

For Lambertian reflectance and $(p_0, q_0) = (0, 0)$ Eq. 2.6 becomes:

$$I(x, y) = \rho \lambda [\cos \sigma + p \cos \tau \sin \sigma + q \sin \tau \sin \sigma] \quad (2.7)$$

where ρ is the albedo of the surface, λ is the strength of the illuminant at the surface, τ is the tilt of the illuminant and σ is the slant.

Eq. 2.7 can be written as

$$I(x, y) = k_1 + p k_2 + q k_3 \quad (2.8)$$

where $k_1 = \cos \sigma$, $k_2 = \cos \tau \cos \sigma$ and $k_3 = \sin \tau \sin \sigma$ and so $L = (k_2, k_3, k_1)$ form the generalized illuminant direction.

Taking Fourier transform of Eq. 2.8 on both sides we get

$$F_I(f, \theta) = H(f, \theta) F_Z(f, \theta) \quad (2.9)$$

where $F_I(f, \theta)$ is the Fourier transform of the image, $F_Z(f, \theta)$ is the fourier spectrum of the surface and $H(f, \theta)$ is a linear transfer function which relates the Fourier transform of the

image to that of the surface.

$$H(f, \theta) = 2\pi f e^{i\pi/2} [k_2 \cos \theta + k_3 \sin \theta] \quad (2.10)$$

Hence given the linear reflectance function, the surface shape can be estimated as follows:

$$F_Z(f, \theta) = H^{-1}(f, \theta) F_I(f, \theta) \quad (2.11)$$

The depth map can be obtained by taking the inverse Fourier transform of eq. 2.11.

2.2.5 Tsai and Shah's method

Tsai and Shah [33] proposed a method which forms a linear approximation of the reflectance map in terms of $Z(x, y)$ instead of p and q as in Pentland's method. Starting with Horn's equation for Lambertian surfaces as in eq. 2.1, p and q are approximated using the forward difference formula as follows

$$p = \frac{\partial Z}{\partial x} = Z(x, y) - Z(x - 1, y) \quad (2.12)$$

$$q = \frac{\partial Z}{\partial y} = Z(x, y) - Z(x, y - 1) \quad (2.13)$$

Eq. 2.1 can now be written as

$$0 = f(I(x, y), Z(x, y), Z(x - 1, y), Z(x, y - 1)) \quad (2.14)$$

$$= I(x, y) - R(Z(x, y) - Z(x - 1, y), Z(x, y) - Z(x, y - 1)) \quad (2.15)$$

Taking a Taylor series expansion for a fixed point (x, y) in an image I about a given depth

map Z^{n-1}

$$\begin{aligned}
0 &= f(I(x, y), Z(x, y), Z(x-1, y), Z(x, y-1)) \\
&\approx f(I(x, y), Z^{n-1}(x, y), Z^{n-1}(x-1, y), Z^{n-1}(x, y-1)) \\
&\quad + (Z(x, y) - Z^{n-1}(x, y)) \frac{\partial f(I(x, y), Z^{n-1}(x, y), Z^{n-1}(x-1, y), Z^{n-1}(x, y-1))}{\partial Z(x, y)} \\
&\quad + (Z(x-1, y) - Z^{n-1}(x-1, y)) \frac{\partial f(I(x, y), Z^{n-1}(x, y), Z^{n-1}(x-1, y), Z^{n-1}(x, y-1))}{\partial Z(x-1, y)} \\
&\quad + (Z(x, y-1) - Z^{n-1}(x, y-1)) \frac{\partial f(I(x, y), Z^{n-1}(x, y), Z^{n-1}(x-1, y), Z^{n-1}(x, y-1))}{\partial Z(x, y-1)}
\end{aligned} \tag{2.16}$$

For an $N \times N$ image there will be N^2 such equations which is solved using Jacobi iterative method. Eq. 2.16 is reduced to a simple form as shown below:

$$Z^n(x, y) = Z^{n-1}(x, y) + \frac{-f(Z^{n-1}(x, y))}{\frac{df(Z^{n-1}(x, y))}{dZ(x, y)}} \tag{2.17}$$

Assuming the initial estimate $Z^0(x, y) = 0$ for all pixels, the depth map can be iteratively solved using eq. 2.17.

2.3 Shape from Shadow

Most of the shape from shading techniques so far are based on assumptions like Lambertian reflectance and point source lighting with no interreflections. But real world scenes are characterized by complex diffuse lighting conditions with interreflections and classic SfS algorithms fail to generalize under such scenarios. Any cast shadows are usually ignored or explicitly removed from the image. But shadows serve as strong depth cues and should be incorporated into depth inference and SfS algorithms. The psychophysical aspects of shadows as depth cues have been studied in [34–36]. Many past approaches to extract shape from shadows used multiple images of the same scene under single point source lighting of

known directions [37,38]. The main downside to this is that the cast shadows created by point source lighting are very well defined with explicit contours whereas the shadows formed under diffuse lighting are not completely darkened regions and can have varying intensity which are a result of various interreflections caused due to other objects in the scene (such as walls, ceiling, etc). Hence making use of shadow cues under diffuse lighting to infer depth can be complicated. Shape from shading under diffuse lighting was first explored by Langer and Zucker [39] and an improved version of their algorithm was presented in [40].

In [4] Potetz and Lee study the statistical correlations between intensity and range images and they make an observation that shadowing results in a direct correlation between the brightness of a pixel and its nearness to the observer which they call the 'da Vinci correlation', as da Vinci used this principle to depict 3D illusions in paintings. We can see this effect in everyday objects with crevices and concavities like the folds of a cloth, piles of objects, foliage, etc where farther regions are darker and most likely the ones to be shadowed regions or concavities. They also observed the effect of shadow cues on depth inference in both urban and rural scenes which showed that shadows contribute significantly more than just shading alone in natural scenes.

2.4 Statistical approaches to depth inference

The difference between probabilistic and statistical approaches is that the former explicitly models the posterior probability distribution $p(Z|I)$ whereas the latter learns this model from the statistics of images. Although most shape inference algorithms are based on inverting physics based image formation models, there have been some exceptions to this. Many computer vision problems can be modeled using large probabilistic graphical models. These models work by factoring large probability distributions into simpler ones. In the recent past, probabilistic inference algorithms like loopy Belief Propagation (BP) have been applied successfully to real-world computer vision problems. But the main drawback

of BP is that the run time gets prohibitively slow and grows exponentially in the size of the largest clique. In cite [41], Potetz proposed a computationally efficient Belief Propagation algorithm. Expectation Propagation (EP), proposed by Minka [42] is a generalization of BP which works by approximating the distributions to a family of analytically simpler ones. In cite [5], the authors proposed an efficient Expectation Propagation algorithm for Shape from Shading by exploiting the second order statistics of natural images and range images. There are some approaches which seek to infer the abstract of the scene or the overall 3D structure without zooming in to the finer 3D detail of local objects/surfaces. In [17], Torralba and Oliva seek to infer the absolute depth of the scene by measuring the energy of wavelet responses at different locations within an image (both artificial and natural) and also the correlations between the wavelet responses which form their feature set v . PCA is used to reduce the dimensionality of the feature set and a mixture of gaussians and EM algorithm is used to learn the joint probability distribution $f(D, v)$ where D is the absolute depth and v is the feature vector. The conditional expected value $E(D|v)$ is obtained by using Bayes rule and a prior model of $p(v)$. Torralba and Oliva in [18] present a model to infer the 3D gist of the scene called the "spatial envelope" described by perceptual dimensions such as naturalness, openness, roughness, expansion, etc which can be estimated using spectral and coarsely localized information.

Another notable approach for creating a 3D model from a single photo is by Derek Hoiem [28,43]. Their focus is on creating models of outdoor images which can be divided into regions belonging to the ground, surfaces that stick out of the ground and the sky. Surfaces that stick out of the ground are further subdivided into planar surfaces facing left, right or towards the camera, porous (like foliage, mesh of wires, etc) or non-porous surfaces (tree trunks, people, etc). Their algorithm starts by first segmenting the image into homogenous regions called "superpixels" and superpixels into "constellations" which are labeled based on color, texture, location and perspective based cues. The labels are learned from a set of training images using logistic regression Adaboost using decision trees. These labels are then used

to create visually pleasing 3D models. An extension to this approach to handle occluding foreground objects is proposed in [29]. Andrew Ng. and his group worked on the same problem to create visually pleasing 3D models from single images [30,31]. They compiled a database of high resolution color images and low resolution range images which is described in the next chapter. Their approach starts by dividing each image into patches and collecting texture-based statistics. These cues along with other monocular cues are incorporated into a joint gaussian multiscale MRF model. They use part of their database for training and the rest for testing purposes.

Chapter 3

Datasets of Single Objects with Ground-Truth Depth

3.1 Introduction

Most algorithms developed thus far for depth inference particularly SfS are tested mostly on synthetic images. High resolution images and range scans are very essential to study the joint statistics between 2D images and 3D structures, which can be used for depth inference and other related vision applications like object recognition, Content Based Image Retrieval (CBIR) and computer graphics applications. Currently there are few other datasets consisting of natural range images of cluttered scenes. Potetz and Lee [4] collected a database of co-registered natural images and range data using a Reigl LMS-Z360 range scanner. They studied the joint statistical relationships between image and range data [4] and used the statistics for depth inference [21]. Another such database was collected by Dale Purves [15] but his research is mainly on psychophysical aspects of range statistics alone. Both of these databases contain images of cluttered natural outdoor scenes with the camera oriented parallel to the ground. The statistics of such scenes are highly regular due to the ground plane and the tendency for any structures built on the ground to be vertically oriented. Depth

inference algorithms in such environments can rely heavily on those regularities. My research is focused on depth inference of single objects. Also these databases have no information on photometric properties for each surface - images are obtained under a single lighting condition and all surfaces have unknown albedo and reflectance functions. The third database was collected by Andrew Ng [30]. Their database has high resolution 2270×1704 color images but only low resolution 55×305 range images which they collected using a custom built laser scanner. They also report some alignment errors between the range and color data of about 2 depth patches. Their work has focused on depth inference from a single image using both monocular and stereo cues for applications in obstacle detection in autonomous cars [30] and also creating visually pleasing 3D models or 3D fly-throughs [31]. The Cornell Activity Datasets (CAD) [44,45] were collected using Microsoft Kinect. The Microsoft Kinect sensor consists of an RGB camera and an infrared structured light depth sensor. These datasets were compiled for human activity detection and recognition algorithms. The CAD datasets contain scenes comprising of unstructured environments like a living room, office, kitchen, bathroom, etc with a subject performing an activity like chopping an onion, making cereal, brushing teeth, etc. The CAD 60 and CAD 120 datasets contain 60 and 120 240×320 p RGB-D videos respectively of people performing daily activities in daily environments. Another range dataset is one by Chene et. al. [46] which comprises of co-registered RGB and range image pairs of natural images taken in a wooded area comprising of dense foliage.

Other available range datasets are the USF dataset which contains 400 co-registered grey scale intensity and range data and the Middlebury stereo dataset which was compiled to mainly benchmark stereo algorithms. Stereo datasets will be discussed in more detail in section 3.2.4.

We have acquired two new databases of range scans and color images of single objects viewed from a fixed distance. These databases capture the full 3D shape of individual objects from multiple viewing angles.

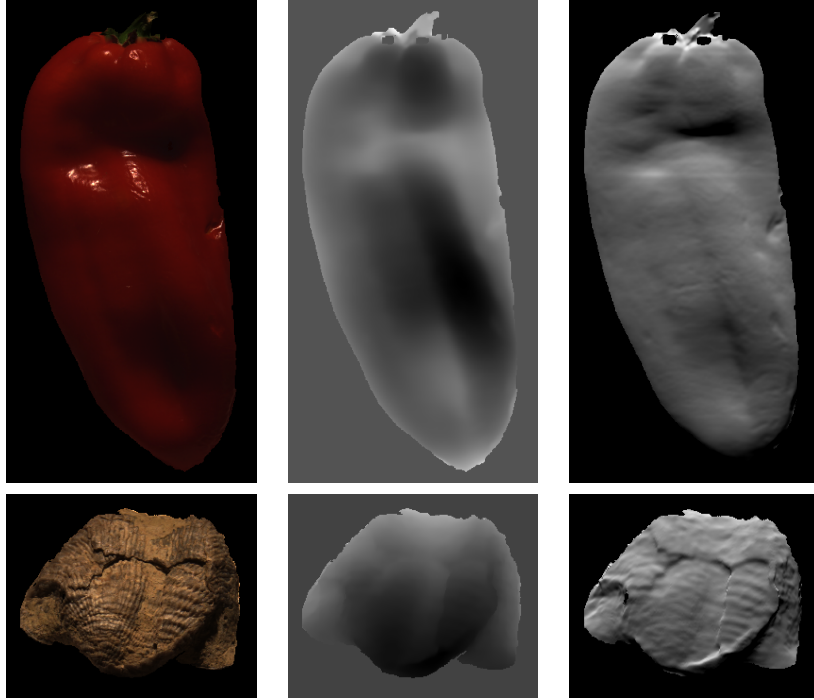


Figure 3.1: Example objects from the SORH database, with color images in the first column and corresponding range images in the second column (with more distant pixels depicted as brighter). The third column shows an artificial rendering of the objects using a Lambertian reflectance model with a similar lighting direction as the color images.

3.2 Single Object Range and HDR (SORH) Database

We collected a database of 26 high resolution range images with co-registered High Dynamic Range (HDR) color images for a variety of natural and man-made objects. Example objects in the database include a pepper, a banana, a muffin, PVC plumbing fixtures, small toys and miscellaneous objects. Fig. 3.1 shows example images from the SORH database.

3.2.1 Acquisition Process of HDR Images and Range Scans

Two Nikon D90 SLR cameras were used to capture high resolution images of single objects under four different exposure settings of $\frac{1}{30}$, $\frac{1}{3}$, 3 and 30s. We used two cameras to obtain left and right stereo pairs of images and range scans. The images were 4288×2848 and were taken in NEF (RAW format for Nikon SLR cameras) format in order to avoid any compression

related artifacts. This process was repeated for nine separate point light sources of known directions. Fig. 3.2 illustrates the approximate positions of the camera, object and the light sources. After all color images were completed, the range scans were obtained by recording videos of each object as a high power green line laser was swept across the object surface. All videos were recorded in 1280x720p resolution using the Nikon D90 cameras used to acquire the color images, thus ensuring that the range and color images were perfectly co-registered. Scans were taken in a completely dark environment to provide maximal contrast for the line laser. Objects that were specular or shiny were lightly spray-painted before scanning with a matte white paint to reduce interreflections of the laser on the object surface. Objects having very dark regions tend to absorb the laser beam falling on them. Such objects were spray-painted as well. Fuzzy/furry objects are not ideal targets for laser scanning and were avoided. The laser was mounted on a 15 rpm motor. For each range image, the laser was swept across the object four times: with the laser positioned from the above-left, from the below-left, from the above-right, and from the below-right of the object. This ensured that nearly all pixels on the object surface were scanned by the laser at least once. The stitched videos of the four laser scans were fed into the DAVID laser scanning software [47,48] which uses optical triangulation to calculate the 3D surface/location of the target object.

Taking four different frontal scans from above and below were not enough to capture every nook and cranny of the objects, especially the boundaries. Hence the process was repeated to obtain scans of two additional views of the object: by rotating it left and right with respect to the frontal view along the y-axis. Care was taken not to introduce any horizontal translations to the object position. The resulting 3 range scans were then fused using DAVID shape fusion to get a rotatable 3D mesh. The fusion software uses surface registration/alignment which works by searching for the relative pose between two scans that maximizes contact area. The aligned scans were then fused into a single triangle mesh. The resulting 3D meshes were stored in .obj format for point cloud data. The image in fig. 3.3 shows the trial setup used for data acquisition with just one camera and a handheld line

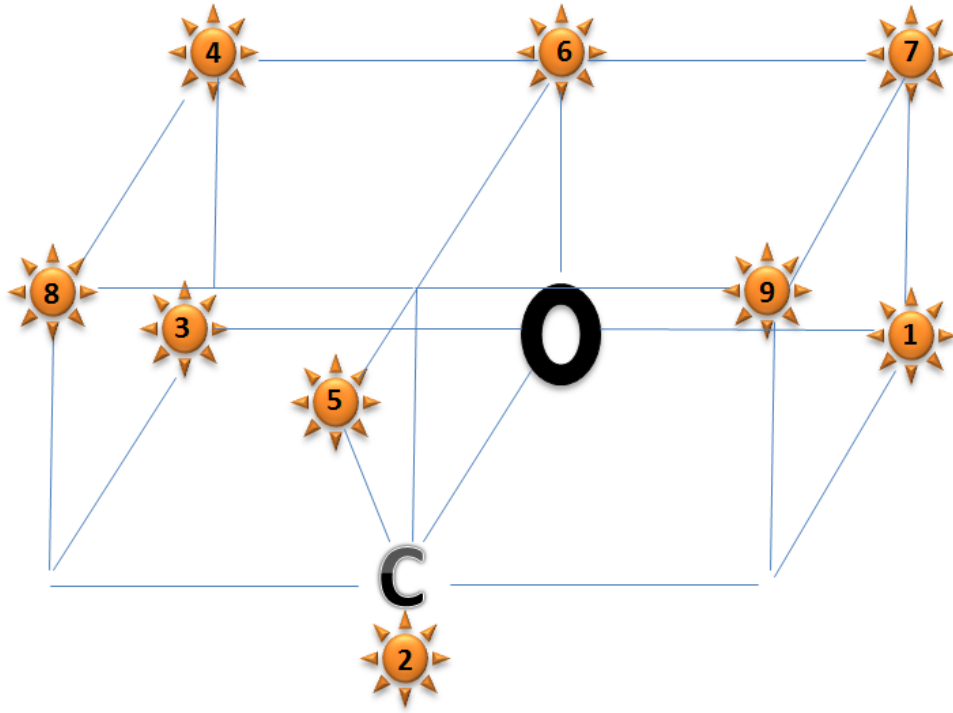


Figure 3.2: Approximate positions of the camera (C), object (O) and the nine light sources.

laser during the initial stages of my project. Once the process of scanning was perfected, two cameras were mounted on a stereo rig and a motorized laser was used. Fig. 3.2 provides an illustration of the final data acquisition setup.

When collecting scans, strange grid artifacts were observed in the final range images. The videos captured by Nikon D90 have a stair-stepping/aliasing effect which was causing the grid artifacts. It seems that the camera actually captures a 800p video and scales it down to 720p. We used the D90 rescaler plugin for Final-Cut software by Mattias Sandstrom and wrote our own Matlab version to get a 1280x800p video which removed the grid artifacts.

3.2.2 High Dynamic Range (HDR) imaging

Cameras (digital/film) seldom capture the true radiance values in the scene. There is a non-linear mapping between the irradiance captured by the imaging sensor and the final pixel intensities that are stored. This non-linear mapping is a result of the various stages in



Figure 3.3: Intial experimental setup for laser range scanning.

the imaging pipeline: noise due to the electronics, resolution of analog to digital conversion, signal processing and compression. This results in saturation points where scene intensity values beyond a certain value are mapped to the same maximum/minimum value. Hence, cameras are unable to capture the entire dynamic range of real scenes. All of this poses a problem for computer vision problems like photometric stereo, shape from shading, etc which work by assuming that pixel intensities are proportional to the true radiance values. The technique of Debevec and Malik [49] was used to recover the camera response function (non-linear mapping), using photographs of the same object taken under multiple exposures and combined the images into a single High Dynamic Range (HDR) radiance map. The

details of the method are reproduced from [49].

The non-linear mapping from the scene irradiance to the pixel intensities can be written as:

$$Z_{ij} = f(E_i \Delta t_j) \quad (3.1)$$

where E_i is the true irradiance at pixel i , Z_{ij} is the recorded pixel intensity at pixel i and j indexes over exposure times Δt_j . Z_{ij} and Δt_j are known values while E_i and f are the unknowns which are to be estimated. The inverse response is given by:

$$f^{-1}(Z_{ij}) = E_i \Delta t_j \quad (3.2)$$

Taking logarithms on both sides, we have:

$$\ln f^{-1}(Z_{ij}) = \ln E_i + \ln \Delta t_j \quad (3.3)$$

Defining $g = \ln f^{-1}$,

$$g(Z_{ij}) = \ln E_i + \ln \Delta t_j \quad (3.4)$$

The unknowns E_i and g are estimated by minimizing the following least squares objective function:

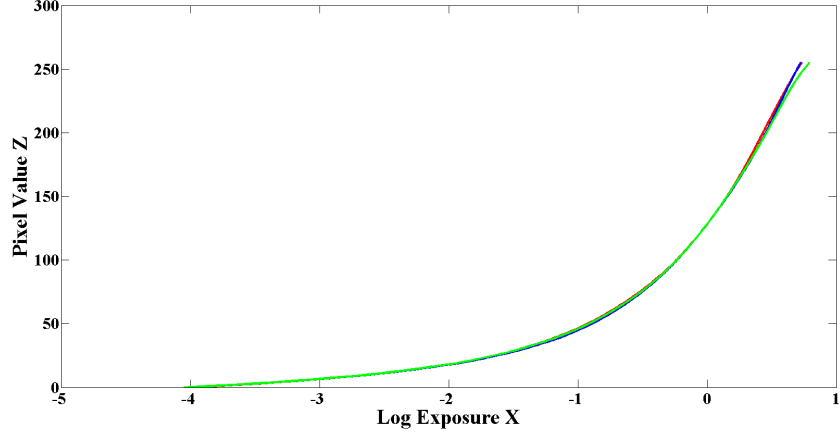


Figure 3.4: The retrieved response functions for the red, green and blue channels of the Nikon D90 DSLR camera.

$$O = \sum_{i=1}^N \sum_{j=1}^P w(Z_{ij}) [g(Z_{ij}) - \ln E_i - \ln \Delta t_j]^2 + \lambda \sum_{z=Z_{min}+1}^{Z_{max}-1} w(z) g''(z)^2 \quad (3.5)$$

where Z_{min} and Z_{max} are the minimum and maximum pixel intensity values, N is the number of pixel locations and P is the number of photographs/exposures. The second term in the equation is a regularization term to ensure that the function g is smooth. $w(z)$ is a hat function of the form:

$$w(z) = \begin{cases} z - Z_{min} & \text{for } z \leq \frac{1}{2}(Z_{min} + Z_{max}) \\ Z_{max} - z & \text{for } z > \frac{1}{2}(Z_{min} + Z_{max}) \end{cases}$$

$w(z)$ emphasizes pixel values that fall near the mid of the pixel range and decays as it reaches Z_{min} and Z_{max} , where $g(z)$ is less smooth and fits the data poorly.

To recover the camera response curve, multiple photographs of the same scene assembled especially for this purpose were taken. The scene contained very dark and shiny objects which would result in very dark and over-saturated pixels respectively. The photographs

were taken at 18 exposure times ranging from $\frac{1}{4000}$ s to 30s with the former being under exposed and the latter over exposed. The camera settings like the shutter speed, aperture, f-value, ISO etc can be extracted from the EXIF tags of the images. To make sure that the system of linear equations arising out of equation 3.5 is sufficiently overdetermined, the number of pixels N is chosen such that $N > \frac{(Z_{max}-Z_{min})}{P-1}$. Based on this, N was set to 60 for $P = 18$ exposure settings and a pixel range of $(Z_{max} - Z_{min}) = 255$. Fig. 3.4 shows the camera response functions recovered for the red, green and blue channels of the Nikon D90 DSLR camera.

After recovering g , the high dynamic range radiance values are computed by a weighted combination of the images taken under multiple exposure settings as follows:

$$\ln E_i = \frac{\sum_{j=1}^P w(Z_{ij})(g(Z_{ij}) - \ln \Delta t_j)}{\sum_{j=1}^P w(Z_{ij})} \quad (3.6)$$

Each object in the database was photographed under four exposure settings $(\frac{1}{30}, \frac{1}{3}, 3, 30)$ which were then combined using the pre-computed response curve to get a HDR radiance map.

Digital cameras have a feature called automatic exposure bracketing which takes under-exposed and over-exposed photographs in addition to an optimally exposed photograph. These are then blended together to obtain a higher quality photograph which has a greater dynamic range. Most smartphones today have cameras that can take HDR images. They use exposure bracketing and combine 3 images to get a single HDR image which is then tonemapped and displayed. HDR found in off the shelf cameras relies on calibration curves that are estimated one time for all cameras of the same model and hardcoded into each unit. Tonemapping is a technique which is done to map the high dynamic range radiance values to a lower dynamic range which can be displayable on screens of computers and mobile devices while still preserving the details of HDR images. While tonemapping generates an attractive

image, we are only interested in estimating true scene radiance for our research.

3.2.3 Photometric Stereo

In addition to the range scans, the SORH database has photometric stereo imagery comprising of nine calibrated color images taken under point light sources of known direction.

One of the publicly available datasets for photometric stereo is by Alldrin, Zickler and Kriegman [50]. Their dataset consists of HDR images of an apple and two gourds. 102 images were taken under different point light sources for the apple and 112 images were taken for the gourd respectively. The dataset does not include groundtruth range data. Also the objects have simple overall surface shape without any significant texture variations.

Another dataset is by Hertzman and Seitz [51,52]. Their main line of research is example-based photometric stereo, in which they make use of reference objects of similar material and known shape and then match points on the target object to the points on the reference object by making use of the orientation-consistency cue, which states that two points with the same surface orientation reflects the same light towards the viewer. Their photometric stereo dataset consists of images of five objects made up of shiny and spatially varying BRDF's, anisotropic material like velvet along with reference spheres used in their algorithm. Again this dataset does not have groundtruth range data.

The main advantage of the SORH database in comparison with the above mentioned datasets, is the presence of high resolution ground truth range data to accurately compare shape estimates of photometric stereo algorithms. The dataset has 26 different objects covering both simple surface shapes and textured surfaces. Apart from the 9 photometric stereo images, videos of the objects were recorded by randomly moving the point light source by hand.

The availability of multiple lighting conditions can be used to accurately estimate shape, surface material reflectance properties at each point as in Goldman et. al's work [53]. Unlike Hertzmann and Seitz's approach [51, 52] which requires reference objects, Goldman et. al's

approach [53] doesn't require any reference objects made up of the same material as the target. They compute both shape and spatially varying BRDFs of objects by modeling the surface as composed of two or more fundamental materials and each pixel as a mixture of these fundamental materials specified by material weight maps:

$$I_{i,p,c} = \sum_m \gamma_{p,m} f_c(\mathbf{n}_p, \mathbf{L}_i, \alpha_m) \quad (3.7)$$

where $I_{i,p,c}$ is the intensity of channel c of pixel p in image i , γ represents the material weight map, the function f_c represents the parameterized lighting model for color channel c with normal \mathbf{n}_p , illuminant direction \mathbf{L}_i and parameter vector α_m where m represents the number of fundamental materials. They used an isotropic Ward reflectance model.

They solve for the surface shape and materials by minimizing the following objective function:

$$Q(\mathbf{n}, \alpha, \gamma) = \sum_{i,p,c} (I_{i,p,c} - (\sum_m \gamma_{p,m} f_c(\mathbf{n}_p, \mathbf{L}_i, \alpha_m)))^2 + Q_{sp}(\alpha) \quad (3.8)$$

The normal maps are initialized using Lambertian photometric stereo thresholding away any specularities. Q_{sp} is a penalty term added to constrain the range of materials/albedos to those observed in the data. The number of fundamental materials is manually selected.

Even after taking measures to capture all nooks and crannies and regions in shadows and concavities, laser range scanning might still miss some spots resulting in gaps of unavailable data. High quality photometric stereo shape estimates could be used to fill such gaps. Like SfS, photometric stereo provides a way to estimate surface normals, which allows us to capture even minute deviations in an object's surface. In theory, the SORH database could be used to improve the ground-truth depth estimates using photometric stereo and use that to improve algorithms that infer shape from single images. The database can be used to

benchmark or train algorithms that estimate shape from photometric stereo, as long as they inferred shape from the low resolution images which is accurate in our database.

3.2.4 Binocular Stereo

A well-known database for binocular stereo is the Middlebury stereo dataset [54]. Their datasets have been compiled over time from 2001 to 2006, comprising of high resolution stereo sequences of 38 complex scenes with pixel-accurate ground-truth disparity data, acquired using structured lighting technique to form unique correspondences [55,56]. Each scene has around 9 color images which were taken at equally spaced view-points translated along the x-axis from left to right taken under multiple illumination conditions and multiple exposures. Each scene has 2 ground-truth disparity maps. The datasets are provided in three resolutions ranging from high to low.

The SORH database also contains binocular stereo imagery. Two Nikon D90 DSLR cameras were used to simultaneously obtain left and right stereo pairs of images and range scans under multiple lighting directions. The Middlebury stereo dataset comprises of complex scenes consisting of multiple occluding objects. The SORH database on the other hand comprises of HDR images of single objects and their corresponding high resolution range imagery. Depth can be calculated from the disparity maps computed from the stereo image pairs using a simple formula:

$$z_i = \frac{fB}{d_i} \tag{3.9}$$

where f is the focal length of the camera in pixels, B is the baseline which is the separation between the two cameras, d_i is the computed disparity at pixel i and z_i is the depth at pixel i . The computed depth can then be compared against the corresponding groundtruth range image. Also the presence of multiple lighting directions can be used to evaluate algorithms

which make use of both photometric and binocular stereo techniques to estimate depth as in [57]. Combining photometric stereo and binocular stereo might potentially result in high quality 3D depth estimates as it makes use of the high resolution details that can be captured by photometric stereo and metric depth information that can be captured by binocular stereo. Metric depth means the exact depth of the point with respect to the camera and the relative depths between objects which cannot be captured by photometric stereo.

The main weakness of the Middlebury stereo dataset is that many of the range images have low depth resolution. The range images are very discretized and only have a few values. For example, the SORH scans have plausible looking point cloud (.obj) files (at least, the front of the .obj files are reasonable). Some of the finest detail might be missing, but they capture the shape and medium scale details very well. In contrast, if we try to encode an object from a Middlebury scan as a point cloud file and rotate around it, most objects would have only a few depth values. There would be staircase artifacts at best, and some objects may appear as completely flat, like cardboard cut-outs. Benchmarking against the Middlebury dataset encourages binocular stereo algorithms to accurately locate each object in space, and to accurately segment one object from another. But it does not encourage the algorithm to estimate the 3D shape of the individual objects with any accuracy.

Other available stereo datasets are the KTTI vision benchmark [58, 59] and the HCI database for stereo and optical flow [60]. The KTTI database was acquired using a station wagon equipped with two pairs of video cameras to record color and grayscale videos as they drove around rural areas and highways. They also used a Velodyne laser scanner and a GPS to capture accurate ground-truth data. Their database consists of 194 training image pairs and 195 test image pairs. The HCI database consists of 11 stereo pairs of video streams taken in outdoor scenes containing both rural and city scapes captured during both day and night. These databases were acquired in complex and uncontrolled natural lighting conditions.



Figure 3.5: Example objects from the Eton Myers database, with color images in the first and third columns and corresponding range images in the second and fourth columns (with more distant pixels depicted as brighter).

3.3 Eton Myers Database

The Eton Myers Collection Virtual Museum is a collection of 74 digitized 3D models of ancient Egyptian art and artifacts from the Eton Myers Collection [61]. The project, funded by the Joint Information Systems Committee (JISC), was performed by the IBM Visual and Spatial Technology Centre (VISTA) using a Minolta Non-Contact 3D Digitiser VIVID-910 and a NextEngine3D HD Scanner (2020i). Final models from these scanners had a depth accuracy of $\pm 0.10\text{mm}$ and $\pm 0.38\text{mm}$ respectively. Full RGB color is recorded at each vertex. Example objects from the virtual museum include pottery, sculptures, a full size sarcophagus, and a mummified hand. The Eton Myers color images were acquired under diffuse lighting conditions as opposed to the Single Object Range and HDR Database which was acquired under point light sources. Moreover the images are not HDR which means

that the joint statistics between intensity and corresponding range images which assume that pixel intensities are proportional to scene irradiance will not be accurate. For each 3D model, range images were computed for 22 different camera positions: eight evenly spaced positions about the equator, six positions at each of $\pm 45^\circ$ latitude, and one at each pole. Thus, 1628 range and color image pairs were constructed from the Eton Myers database. The Eton Myers database differs from the SORH database in the following points:

- The subject matter: The Eton Myers database comprises of museum artifacts whereas the SORH database comprises of small objects ranging from random PVC plumbing, a banana, a pepper, plastic toys and figurines.
- The lighting conditions: The Eton Myers database was acquired under diffuse lighting conditions. SORH was taken under controlled point light sources
- The views: The Eton Myers database was scanned in 22 different camera positions resulting in 3D scans in different views. The SORH objects were scanned in 3 views (one frontal and two additional views by rotating the object left and right with respect to the frontal view).
- SORH contains HDR imagery whereas Eton Myers imagery is not HDR.
- SORH contains photometric stereo and binocular stereo imagery while Eton Myers database does not

Fig. 3.5 shows example images from the Eton Myers database.

Please refer to appendix A for the complete listing of imagery from the SORH dataset.

Chapter 4

The statistics of the 3D shape and appearance of single objects and surfaces

4.1 Introduction

The study of the statistics of natural scenes has been pivotal in the advancement of computer vision, perceptual psychology, and visual neuroscience. Statistical knowledge from databases of natural images [7, 62] has been used to improve image compression algorithms [63], image denoising and inpainting [64–66], image segmentation [67], steganalysis [68], and computer graphics [69, 70]. In neuroscience, scene statistics have been essential in explaining the behavior of the retina [71] and simple [6, 9] and complex cells [11, 72, 73] in V1. The study of the statistics of natural scenes has also progressed beyond images, and has included statistical studies of 3D shape and range images [4, 13, 21, 46], video [74], surface albedo [75], illumination [76], and image segmentations chosen by human subjects [77]. Results from these studies have contributed to improved algorithms for the inference of shape, albedo, and segmentation, and have provided new insights into depth inference in the brain [78].

To date, studies of the statistics of natural scenes have focused on cluttered natural scenes with many objects. Studies of natural scenes have often concluded that occlusion

and the distribution of objects within a scene dominate the statistics of natural images [79] and natural range images [13]. Moreover, our theoretical understanding of natural scenes is heavily rooted in our understanding of occlusion, and models of natural scenes based purely on occlusion are able to account for nearly all observed statistical findings of natural images [14]. Because of this, little is known about the statistics of the appearance or the shape of single objects.

At the same time, single objects play an essential role in our interaction with our environment and our visual needs. The problem of inferring 3D shape of a single object is a task that has received substantial investigation, and has applications in object recognition, robotic grasping, defects inspection, content-based image retrieval (CBIR), and others. Such methods can benefit from a stronger understanding of statistical properties of single objects and their range images.

This chapter investigates into the statistics of the 3D shape and appearance of single objects directly. We begin by presenting theoretical arguments that images of unoccluded surfaces should obey similar correlational statistics as those found in cluttered natural scenes (i.e. a $1/f^2$ 2D power spectrum). However, we argue that range images of unoccluded surfaces should obey rather different statistical trends, which predict a $1/f^4$ 2D power spectrum. We test this hypothesis by compiling two datasets comprising of co-registered color and range image pairs along with 3D mesh information of scenes containing single objects. We also find that the $1/f^4$ scale invariance observed in the surfaces of single objects occurs despite a lack of observable scale invariance in the coarse overall shape of the objects themselves. The results suggest that while range images of single objects exhibit scale invariant properties, whole 3D objects are not fully scale invariant and they can be modeled as simple 3D objects wrapped in scale invariant texture.

4.2 Background

Among the most well-known and well-studied of natural scene statistics is the observation that ensembles of natural images have a power spectrum that obeys a $1/f^\alpha$ power-law, with α close to two [7, 8, 62, 80, 81]. This finding is highly robust; it is observable over a range of more than three octaves [7], and has been observed in a wide variety of categories of natural and urban scenes [82]. The $1/f^2$ power spectrum of images is interesting because it implies that the second-order statistics of natural scenes is *scale invariant*: if an ensemble of images $I(x, y)$ is downsampled, resulting in $I(\sigma x, \sigma y)$, the ensemble will have the same power spectra as the original ensemble.

Several possible reasons for scale invariance have been hypothesized. One hypothesis is that images are comprised of objects that may be viewed from many distances. On one hand, due to projective camera geometry, the appearance of objects change when viewed from close up; photographing an object from far away is not equivalent to scaling an image down. Nevertheless, for sufficient distances or sufficiently flat objects, the effects of camera distance approximate the effects of scaling. Note, though, that if this explanation was the sole cause of scale invariance, it would impose constraints on the distribution of distances from camera to object.

Another hypothesis is that natural objects themselves display features of self-similarity. Physical processes that occur at one scale often occur on other scales as well. Many natural objects, such as plants and trees, animals, rocks, and terrain have been modeled as fractal systems [83].

Ruderman shed light on the origin of scale invariance by examining components of the autocorrelation function of natural scenes [79]. The *difference function* $D(r)$ is defined as the mean squared error between two pixels separated by distance r , averaged over all locations,

angles, and images in the ensemble:

$$D(r) = \mathbf{E}[(I(\vec{x}_0) - I(\vec{x}_1))^2 \mid |\vec{x}_0 - \vec{x}_1| = r] \quad (4.1)$$

Assuming that natural scenes are stationary (i.e. that natural images exhibit the same statistical properties at all locations within the image), a power-spectrum of $1/f^\alpha$ implies that the difference function obeys power-law of the form $C - B/r^{E-\alpha}$, for some constants B and C , and where E is the dimensionality of the signal. In the case of 2D images, $E=2$. In the case that $\alpha=E$, the difference function follows a logarithmic form [79]. If the images in the ensemble have been manually segmented by object boundaries, the difference function can be decomposed into subcomponents. Let $D_{in}(r)$ be the difference function averaged only over pixel pairs such that both \vec{x}_0 and \vec{x}_1 lie inside the same object. Let $D_{out}(r)$ be defined similarly as the difference function over only pixel pairs that lie across different objects. Let $P_{in}(r)$ be the probability that two pixels, separated by distance x , lie within the same object. Then $D(r)$ can be written as

$$D(r) = P_{in}(r)D_{in}(r) + (1 - P_{in}(r))D_{out}(r) \quad (4.2)$$

Using manually segmented images, Ruderman showed that $D_{in}(r)$ and $D_{out}(r)$ were relatively flat with respect to r , and that the form of $D(r)$ could be attributed almost entirely to $P_{in}(r)$.

This observation provided evidence that the scale invariance observed in the power spectrum of natural images was caused primarily by occlusion. More specifically, if each segment in one of Ruderman's segmented images were replaced by a random uniform color, the image would have a similar power spectrum. Ensembles of such images would be scale invariant.

This inspired the *collage* or *dead-leaves* models of natural images, where images were modeled by randomly-placed opaque, overlapping solid objects [14, 79, 84]. Figure 4.1 shows an example collage model image. More formally, the collage model consists of a Poisson point process in 2D space, where each point specifies the center of an opaque object of random size

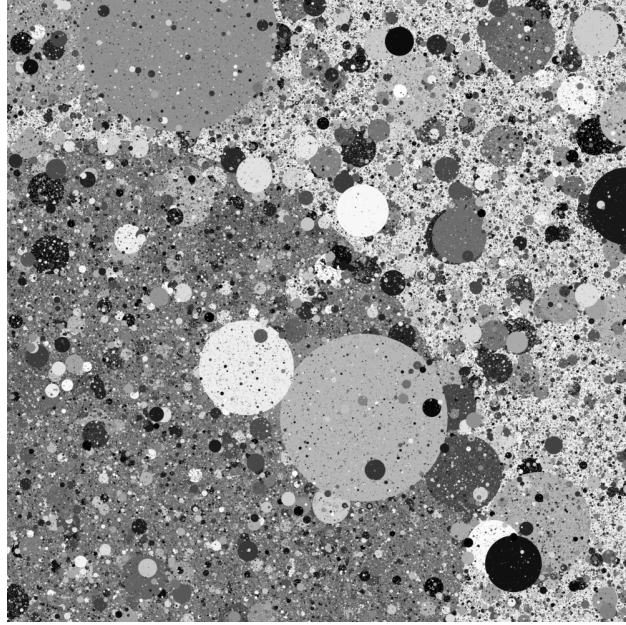


Figure 4.1: Example collage model image with $1/r^3$ distribution of object radii and $1/f^2$ power spectrum.

and color [85]. For any image comprised of uniformly colored regions, the visible area A of each region must obey a distribution of $1/A$ in order to achieve a $1/f^2$ power spectrum [85]. Collage models can be generated by placing a series of opaque 2D objects at random locations within the image, one on top of the others, until the entire image is covered. This process is why the collage model was known as the “dead leaves” model when it was originally developed for use in the material sciences [86]. In order for the visible regions of the 2D “leaves” to obey a $1/A$ distribution, the radii of the 2D objects (which may be partially obscured) must be drawn from a $1/r^3$ distribution [79].

If a family of images is fully scale invariant, then its visually-distinct regions must obey a $1/A$ distribution. More specifically, suppose that an ensemble of images can be segmented into non-overlapping regions, and the choice of segmentation does not depend on scale (so that scaling an image and then segmenting produces the same result as segmenting and then scaling the segmentation). Let $p(A)$ be the distribution of region areas in the image ensemble, and let $p_d(A_d)$ be the distribution of region areas after downsampling the image:

$f(\sigma x, \sigma y)$. Since $A_d = A/\sigma^2$, by change of variables we have that $p_d(A_d) = \sigma^2 p(\sigma^2 A_d)$. If the ensemble is scale invariant, then $p_d(A) = p(A)$, and so $p(A) \propto 1/A$. In the collage model, the distribution of visible image regions is directly related to P_{in} , and therefore to the power spectrum. Steeper power laws over A result in power spectra with more shallow power laws over f .

It has been argued that the power spectrum of natural images has not been measured accurately enough or over a broad enough range of frequencies to reliably draw conclusions about the size distribution of objects in natural scenes [87, 88]. Later, the distribution of visible region sizes was measured directly when it was shown that the sizes of image regions from image segmentations drawn by human subjects produce a $1/A$ Pareto distribution, matching the predictions of scale invariance and the collage model [77]. Interestingly, the sizes of posterized gray-level regions within natural images has been shown to obey a $1/A^2$ Pareto distribution [89]. This discrepancy may be due to the fact that identifying separate regions within the posterized image relies on a connected component computation, which requires some notion of connectivity among image pixels (e.x. four nearest neighbors). These notions of connectivity are defined at a specific spatial scale: two like-colored image regions are considered part of the same region if they are separated by not more than the width of one pixel. Downsampling the image and recomputing the connected components would result in a different partitioning. In particular, small regions are likely to be combined after downsampling, thus reducing the number of small regions identified relative to large regions. Recall that the proof above that scale invariance produces a $1/A$ Pareto distribution over visible region areas relied on the assumption that the choice of segmentation does not depend on scale.

It should be noted that the visual regions of segmentations (including Ruderman's and others') are not limited to individual objects, and may include object parts [77, 79]. It is unclear if the $1/A$ law of visual region areas also applies to the visible portion of physically distinct 3D objects.

One other criticism of the collage model is that single images drawn from the collage model exhibit $1/f^\alpha$ power spectra, whereas single natural images often deviate significantly from a power law power spectrum [81, 90]. It is possible that this could be addressed in the collage model by introducing dependencies between the radii of nearby shapes, or latent variables that capture properties specific to individual images. Still, it is possible that the collage model imposes unnecessarily strict constraints on the distribution of object sizes in natural scenes. It is possible for individual images to contain distributions of object radii that deviate from the $1/r^3$ distribution, and yet still achieve a $1/f^2$ power spectrum of the image ensemble.

Nevertheless, the collage model explains a wide variety of statistical trends observed in natural scenes. In addition to the roughly $1/f^2$ power spectrum of natural images, the collage model also explains the full marginals over the output of various linear filters, the joint distributions over wavelet coefficients, and two-pixel co-occurrence statistics [14].

Also of interest are the statistical properties of natural 3D shapes. A *range image* is an image where the value of each pixel corresponds to the depth of the scene at that point. If the second order statistics of natural images is attributable to occlusion, then we would expect to see similar statistical properties in range images, which share the occlusion boundaries of color images. This has been confirmed using datasets of natural range images acquired by laser range scanners or using structured light [13, 21, 46].

The traditional collage model is strictly a 2D model, where thin overlapping 2D shapes are arranged randomly in 2D space. Extending the model to describe the 3D world would require constraints on both the 3D sizes of objects, the arrangements of objects in space, and the placement of the camera [91]. A more realistic 3D model may require a more detailed account of the arrangement of objects in 3D; rather than floating randomly in space, in real scenes objects are clustered or adjacent, and hierarchical processes that account for the relationship between parts of objects to its whole may have benefits over a 3D Poisson point process [14].

Ruderman’s observation that $D_{in}(r)$ is comparatively constant in natural scenes relative to $P_{in}(r)$ is direct evidence that the covariance structure of natural scenes is dominated by the effects of occlusion. Moreover, the collage model is an image model that ignores all effects other than occlusion, and its success in accounting for a wide variety of statistical trends in natural scene suggests that much of what is known about natural scene statistics can be attributed to occlusion. Additionally, studies in computer graphics have found that the covariance structure of artificially generated images is highly sensitive to scene geometry, but none of the many rendering parameters and choices significantly affected the power spectra [92].

Thus, there is little that is known about statistical trends in the appearance and 3D shape of single objects. So far, a direct study of the statistics of single objects has not been undertaken. And yet single objects play an important role in many visual tasks. In particular, the inference of the 3D shape of single objects is one such task that has applications in object recognition, robotic grasping, CBIR, defects inspection and others. Such algorithms stand to benefit from strong probabilistic priors over appearance and shape. In addition, all computer vision tasks that benefit from a strong understanding of scene priors, such as compression, denoising, inpainting, steganalysis, binocular stereo, image segmentation, and computational photography may benefit from a more detailed understanding of the statistical trends of the appearance and shape of single objects. Although these algorithms often operate over whole, cluttered scenes, stronger statistical priors on the properties of individual objects may improve algorithm performance within object boundaries. Even in circumstances where algorithm performance within objects may play a minor role in sum-squared error metrics often used for performance evaluation, improvements in handling objects may yield a strong gain on subjective performance measures due to the importance that human observers place on objects.

To investigate the statistical properties of range images and 3D shapes of single objects, we compiled two datasets comprising of co-registered color and range image pairs along

with 3D object files of scenes containing single objects (both natural and man-made). The datasets are described in detail in chapter 3. Sections 4.3.1 and 4.3.2 build the theoretical framework of our hypothesis.

4.3 A Scaling Hypothesis for the Range Images of Single Objects

We start by providing a detailed derivation explaining the self similarity of autocorrelation in image ensembles in section 4.3.1. It is shown mathematically that natural image ensembles are scale invariant when their power spectra follow a $1/f^2$ power law. Section 4.3.2 studies the statistics of range images as a consequence of scale invariant properties of whole 3D objects. It is derived that if 3D scenes of natural images are scale invariant, the corresponding range images of unoccluded surfaces are self affine with a $1/f^3$ power law power spectrum averaged over orientation.

4.3.1 A Derivation of the Consequences of Self-Similarity

Let $I_i(\vec{x})$ be an ensemble of images, $1 \leq i \leq N$. It will be helpful later to remain general about the dimensionality of the signal. Let E be the number of dimensions of the signal, where E will be two for 2D images. Thus, \vec{x} is a vector of length E . We will assume that images are stationary, so that image statistics do not depend on location within the image. The autocorrelation function is defined as

$$h[\{I_i(\vec{x})\}](\vec{x}) = \mathbf{E}[I_i(\vec{x}_0)I_i(\vec{x}_1) \mid \vec{x}_0 - \vec{x}_1 = \vec{x}] \quad (4.3)$$

where the expectation is taken over all images i and all locations \vec{x}_0 and \vec{x}_1 separated by an offset of \vec{x} . When the ensemble is unambiguous, we will denote the autocorrelation function simply as $h(\vec{x})$. The autocorrelation function, coupled with the mean, capture the complete

covariance structure of the ensemble.

Let $M[\{I_i(\vec{x})\}]$ be some statistical measure taken over the image ensemble I . M is the expected value of some function of an image, averaged over all images in the ensemble and all spatial offsets within each image. For instance, M may be the mean or autocorrelation of the ensemble. We will say that the measure M is *self-similar* or *self-affine* if for all scalar values σ ,

$$|\sigma|^\eta M[\{I_i(\sigma\vec{x})\}] = M[\{I_i(\vec{x})\}] \quad (4.4)$$

If $\eta=0$, we can say that M is *scale invariant* over I .

Any image statistic may exhibit self-similarity. To begin, however, let us examine the self-similarity of the autocorrelation function h . It is convenient to use spherical coordinates to index h ; let $(r, \vec{\theta})$ represent the same point as \vec{x} in spherical coordinates. Then, if h is self-similar, for some function $b(\vec{\theta}) = h(1, \vec{\theta})$ we can write:

$$|\sigma|^\eta h[\{I_i(\sigma\vec{x})\}](\vec{x}) = h[\{I_i(\vec{x})\}](\vec{x}) \quad (4.5)$$

$$|\sigma|^\eta h(\sigma r, \vec{\theta}) = h(r, \vec{\theta}) \quad (4.6)$$

$$h(\tilde{r}, \vec{\theta}) = \tilde{r}^{-\eta} b(\vec{\theta}) \quad (4.7)$$

where we have substituted $r \rightarrow 1$ and $\sigma \rightarrow \tilde{r}$.

Unfortunately, by the above derivation, perfectly self-similar signals are only realizable in nature by trivial signals such as constant functions. For example, in order to achieve scale invariance ($\eta = 0$), the autocorrelation function must be independent of r . This includes ensembles of uniformly colored solid images (whose autocorrelation is constant), or pure, infinite-bandwidth white noise (each infinitesimal point in continuous Cartesian space is assigned an independent random value). Unfortunately, infinite-bandwidth white noise cannot be realized in nature; any physical medium must have a maximum obtainable frequency. If $\eta < 0$, covariance increases unboundedly as r increases, requiring infinite signal

variance. Likewise, when $\eta > 0$, covariance increases unboundedly as r approaches zero. Thus, the current definition of self-similarity is too restrictive. Typically, this definition is relaxed slightly by requiring only that the ensemble exhibit self-similarity within a certain range of octaves of spatial frequency. To do this, we examine the Fourier transform of the autocorrelation function.

By the Wiener-Khinchin theorem, the Fourier transform of the autocorrelation function of the ensemble I is equal to the power spectrum. Let $H(f, \vec{\theta}) = \mathcal{F}[h(r, \vec{\theta})]$ denote the power spectrum of I . By taking the Fourier transform of equation 4.6, we find that the self-similarity of the autocorrelation function implies that:

$$|\sigma|^\eta \mathcal{F}[h(\sigma r, \vec{\theta})] = \mathcal{F}[h(r, \vec{\theta})] \quad (4.8)$$

$$|\sigma|^\eta |\sigma|^{-E} H(f/\sigma, \vec{\theta}) = H(f, \vec{\theta}) \quad (4.9)$$

$$H(\tilde{f}, \vec{\theta}) = \frac{B(\vec{\theta})}{\tilde{f}^{E-\eta}} \quad (4.10)$$

using the Fourier similarity theorem, and substitutions $f \rightarrow 1$ and $\sigma \rightarrow 1/\tilde{f}$.

We therefore say that ensemble I demonstrates self-similarity of the autocorrelation function h within observable spatial frequencies if the power spectrum obeys the power law $B(\vec{\theta})/f^{E-\eta}$. Images achieve scale invariance when $H(f, \theta) = B(\theta)/f^2$.

It is sometimes convenient to average over orientation while computing the autocorrelation. Let us define

$$h(r) = \mathbf{E}[I_i(\vec{x}_0)I_i(\vec{x}_1) \mid |\vec{x}_0 - \vec{x}_1| = r] \quad (4.11)$$

$$= \int h(r, \vec{\theta}) d\vec{\theta} \quad (4.12)$$

We can derive the Fourier transform of $h(r)$ using the same derivation used for equation 4.10

with E set to 1, so that:

$$\mathcal{F}[h(r)] \equiv H(f) \propto \frac{1}{f^{1-\eta}} \quad (4.13)$$

Note that for two dimensions, we can also prove directly that $H(f, \theta) \propto 1/f^{2-\eta}$ implies that $H(f) \propto 1/f^{1-\eta}$ using the Fourier projection-slice theorem, since

$$\mathcal{F}[h(x, y=0)] \propto \int_{-\infty}^{\infty} \frac{1}{(x^2 + y^2)^{\frac{1}{2}(2-\eta)}} dy \quad (4.14)$$

$$\propto \frac{1}{(x^2)^{\frac{1}{2}(1-\eta)}} = \frac{1}{|x|^{1-\eta}} \quad (4.15)$$

By approximating the inverse Fourier transform of equation 4.10, Ruderman derived the autocorrelation function implied by self-similarity within a finite bandwidth [79], and found that

$$h(r, \vec{\theta}) = \begin{cases} -|C_1| + |C_2|r^\eta, & \text{if } \eta > 0 \\ -\frac{A}{2\pi} \ln(\lambda r), & \text{if } \eta = 0 \\ |C_1| - |C_2|r^{-\eta}, & \text{if } \eta < 0 \end{cases} \quad (4.16)$$

In essence, the constant terms arise because the $f = 0$ component of the power spectrum is no longer restricted when the signal is only constrained to be self-similar within a finite bandwidth.

4.3.2 A Theoretical Justification of $1/f^4$ Power Spectra of Range Within Occlusion Boundaries

As described above, the well-studied finding that images and range images obey a roughly $1/f^2$ power spectrum appears to be caused primarily by the effects of occlusion. The statistical properties of images within image boundaries appears to have been overshadowed by

occlusion effects. Therefore, let us now consider the statistical properties within individual objects directly.

Ruderman found that variance $D_{in}(r)$ within object boundaries was nearly constant in comparison with occlusion-based $P_{in}(r)$ (equation 4.2) [79]. Still, we would expect to see some systematic increase in $D_{in}(r)$ as x increases. From experience, we know that individual objects or other semantic image regions are only rarely colored uniformly; typically image regions contain textures or other local variations. The correlation between pixels within such textures should be expected to diminish with the distance between those pixels. This would produce a power spectrum that decreases monotonically with f .

Let us begin by examining the statistics of range images of the surfaces of single objects, since color images are subject to more complex image formation. In particular, let us consider the consequences that 3D scale invariance of the geometry of natural scenes would have on the statistics of the range images of single objects. Define $b(x, y, z)$ to be a binary, volumetric function that represents occupancy within three-dimensional space: $b(x, y, z) = 1$ for all points in space that are occupied by some opaque object, and $b(x, y, z) = 0$ for all other points. As written in equation 4.4, b is scale invariant if the expected value of any function of b is independent of the scale of b :

$$M(\{b_i(\sigma x, \sigma y, \sigma z)\}) = M(\{b_i(x, y, z)\}) \quad (4.17)$$

Suppose that $z(x, y)$ is a range image of a natural scene, and suppose that we restrict our analysis to a region of z that lies on a single object. For now, let us further restrict our analysis to a region that lies within any self-occlusion contours, so that the analyzed region of z is free of discontinuities. Let us assume that the distance from the camera to the object is considerably larger than the variance in depth along the object surface. This allows us to approximate the perspective projection of pinhole cameras with orthographic projection. We will consider equation 4.17 where the operation M refers to the autocorrelation of the

range image of $b_i(x, y, z)$ when viewed from afar (e.g. outside of the limits of the definition of b_i).

If $z(x, y)$ is the range image corresponding to the scene $b(x, y, z)$, then the scaled scene $b(\sigma x, \sigma y, \sigma z)$ has range image $\frac{1}{\sigma}z(\sigma x, \sigma y)$. Note that the object has scaled along both the x and y axis, and also in depth. By equation 4.3, the autocorrelation of $\frac{1}{\sigma}z(\sigma x, \sigma y)$ can be written as

$$h[\{\frac{1}{\sigma}z_i(\sigma\vec{x})\}](\vec{x}) = \mathbf{E}[\frac{1}{\sigma}z_i(\sigma\vec{x}_0) \cdot \frac{1}{\sigma}z_i(\sigma\vec{x}_1) \mid \vec{x}_0 - \vec{x}_1 = \vec{x}] \quad (4.18)$$

$$= \frac{1}{\sigma^2} \mathbf{E}[z_i(\vec{x}_0) \cdot z_i(\vec{x}_1) \mid \vec{x}_0 - \vec{x}_1 = \sigma\vec{x}] \quad (4.19)$$

$$= \frac{1}{\sigma^2} h[\{z_i(\vec{x})\}](\sigma\vec{x}) \quad (4.20)$$

If equation 4.17 holds, then the autocorrelation of $\frac{1}{\sigma}z(\sigma x, \sigma y)$, averaged over an ensemble of objects, should match the autocorrelation of $z(x, y)$:

$$h(\vec{x}) = \frac{1}{\sigma^2} h(\sigma\vec{x}) \quad (4.21)$$

Thus, if natural scenes are scale invariant, then the range images of unoccluded surfaces are not scale invariant. Instead, they are self-affine with $\eta = -2$ in equation 4.4. In section 4.3.1, we derived the consequences of self-affinity on the power spectrum. Recalling equations 4.10 and 4.13, self-affinity with $\eta = -2$ is equivalent to observing either of the following:

$$\mathcal{F}[h_{cnts}(r, \theta)] = H_{cnts}(f, \theta) = \frac{B(\theta)}{f^{E-\eta}} = \frac{B(\theta)}{f^4} \quad (4.22)$$

$$\mathcal{F}[h_{cnts}(r)] = H_{cnts}(f) \propto \frac{1}{f^{1-\eta}} = \frac{1}{f^3} \quad (4.23)$$

Note that this result does not hold if the object contains self-occlusions. Natural 3D objects often exhibit non-trivial topology (i.e. holes or handles), or complex shape features such as outcroppings, tendrils, folds, and cusps along surfaces. These features can all produce

self-occlusion. Suppose that the object does contain self-occlusions, and let $S(x, y)$ be a binary image describing the locations of occlusion boundaries, so that $S(x, y) = 1$ for all points (x, y) that lie along occlusion contours within the object. The scaled object $b(\sigma x, \sigma y, \sigma z)$ has range image $\frac{1}{\sigma}z(\sigma x, \sigma y)$, and it has occlusion contours $S(\sigma x, \sigma y)$. If 3D objects are scale invariant, then $S(\sigma x, \sigma y)$ would have similar statistical properties as $S(x, y)$, which implies that S would have a power spectrum of $1/f^2$. Similar to Ruderman's $P_{in}(r)$ (equation 4.1), we can define $P_{cnts}(r)$ to be the probability that two points in the image, separated by a distance of x pixels, are not separated by an occlusion contour. More specifically, given \vec{x}_0 and \vec{x}_1 such that $|\vec{x}_0 - \vec{x}_1| = r$, if a straight line is drawn on the image from \vec{x}_0 to \vec{x}_1 , if that line crosses an occlusion contour then we will consider the two points to be separated by the occlusion contour. From equation 4.16, we expect $P_{cnts}(r)$ to drop off logarithmically with respect to r . As in equation 4.2, the autocorrelation function can be decomposed as:

$$h(r) = P_{cnts}(r)h_{cnts}(r) + (1 - P_{cnts}(r))h_{occ}(r) \quad (4.24)$$

where $h_{cnts}(r)$ is the autocorrelation measured over only those point pairs \vec{x}_0 and \vec{x}_1 that are not separated by an occlusion contour, and $h_{occ}(r)$ is the autocorrelation measured over only those points that are separated by occlusion.

Strictly speaking, when occlusion edges are present, the resulting range images would neither be self similar nor scale invariant, since equation 4.24 does not obey a power law. Instead, a range image may be thought of as the combination of two self-similar signals: one signal that is scale-invariant ($\eta = 0$) which is based on occlusion, and another signal that is self-similar ($\eta = -2$) which is based on the scale-invariant properties of the 3D shapes of ensembles of single objects. Although the autocorrelation $h(r)$ of range images with occlusion may not obey a power law exactly, it has been observed that when two self-similar signals are combined, the resulting signal often has a power spectrum that is closely approximated by a power-law [93, 94]. In particular, if signal $f_1(\vec{x})$ is self-similar with power spectrum

$1/f^{\alpha_1}$ and independent signal $f_2(\vec{x})$ is self-similar with power spectrum $1/f^{\alpha_2}$, and if $\alpha_1 < \alpha_2$ by a sufficient margin, then the power spectrum of the sum $f_1 + f_2$ can often be coarsely approximated by a power law $1/f^{\alpha_1}$, especially in the higher frequencies. Intuitively, the reason is that in the higher frequencies, f_2 is substantially weaker than f_1 . Thus, for scenes with occlusion, the hypothesis that 3D scenes are scale invariant would predict that the power spectrum of range images should be closely approximated by a $1/f^2$ power law.

For color images, we should expect a different outcome than for range images. Pixel intensity is governed by several factors, including the surface albedo at each point, the surface shading (given by the angle between the surface normal and incoming light), and the intensity of incoming light at each point (including shadowing and ambient occlusion). Consider a simple Lambertian model of image formation, where an image can be decomposed as

$$i(x, y) \propto l(x, y)\rho(x, y) \cos(\phi(x, y)) \quad (4.25)$$

where l is the light intensity at point (x, y) , ρ is the surface albedo, and ϕ is the angle from the light source to the surface normal.

As shown by Pentland, the first order Taylor series approximation of $\cos \phi$ (which is a dominant term when the lighting is oblique) is the derivative of depth along the direction of the incoming light source [32]. If ensembles of $z(x, y)$ have a power spectrum $H_\phi(f, \theta)$ of $B(\theta)/f^4$, then $\cos \phi(x, y)$ will have a power spectrum proportional to $B(\theta) \cos^2(\psi)/f^2$, where ψ is the incoming light direction. In other words, $\cos \phi$ will not be self-similar, but cross-sections of $\cos \phi$ along the direction of incoming light will exhibit $\eta=0$ scale invariance, while cross-sections perpendicular to the incoming light direction will exhibit $\eta=-2$ self-similarity.

Albedo ρ is not determined by the scene geometry, and should instead be expected to scale along with the scene. We might extend our volumetric model $b_i(x, y, z)$ to include albedo at each voxel. As the scale σ of $b_i(\sigma x, \sigma y, \sigma z)$ changes, the resulting albedo map

becomes $\rho(\sigma x, \sigma y)$. This results in $\eta=0$ scale invariance, and a power spectrum $H_\rho(f, \theta)$ of $B(\theta)/f^2$. If the light sources are expected to scale with the scene $b_i(x, y, z)$, then a similar argument may be made for light intensity $l(x, y)$. As mentioned above, when two signals with different scaling rates η are combined, the result is typically dominated by the signal whose power drops more slowly as frequency increases. Thus, we might expect that color images should exhibit $\eta=0$ scaling, even within the boundaries of occlusion contours, as long as the appearance of the surfaces in the ensemble are not too heavily dominated by shading effects.

4.4 Scaling properties of unoccluded surfaces

In section 4.3.2, we hypothesized that within the confines of occlusion contours, range images of single surfaces should demonstrate self-similarity with $\eta = -2$, rather than the scale invariance ($\eta=0$) observed for cluttered natural images and range images. Recall from eq. 4.13 that a scaling law of $\eta=-2$ is equivalent to $\mathcal{F}[h_{cnts}(r)] = H_{cnts}(f) \propto \frac{1}{f^3}$, and that a scaling law of $\eta=0$ is equivalent to $H_{cnts}(f) \propto \frac{1}{f}$. In this section, we measure $h_{cnts}(r)$ directly using the SORH and Eton Myers datasets. For each pixel in each range image, a line of fixed length at an angle θ is drawn across the image. If this line crosses an occlusion contour, the line is discarded; otherwise, we compute the power spectrum of the vector of pixels that fall on that line. For the Eton Myers images “occlusion” is defined based on the object files which are readily available. For two adjacent pixels in the range image, they are defined to be separated by an occlusion contour if the shortest path between those points on the surface of the object file traversed a back-facing facet. For the SORH database we defined an occlusion contour based on the difference in range values between pixels. Two adjacent pixels are said to be separated by an occlusion contour if the difference between their range values exceeds a certain threshold. Power spectra are calculated for vectors of pixels obtained at 36 different angles each separated by 5 degree intervals. A Hanning window is applied to

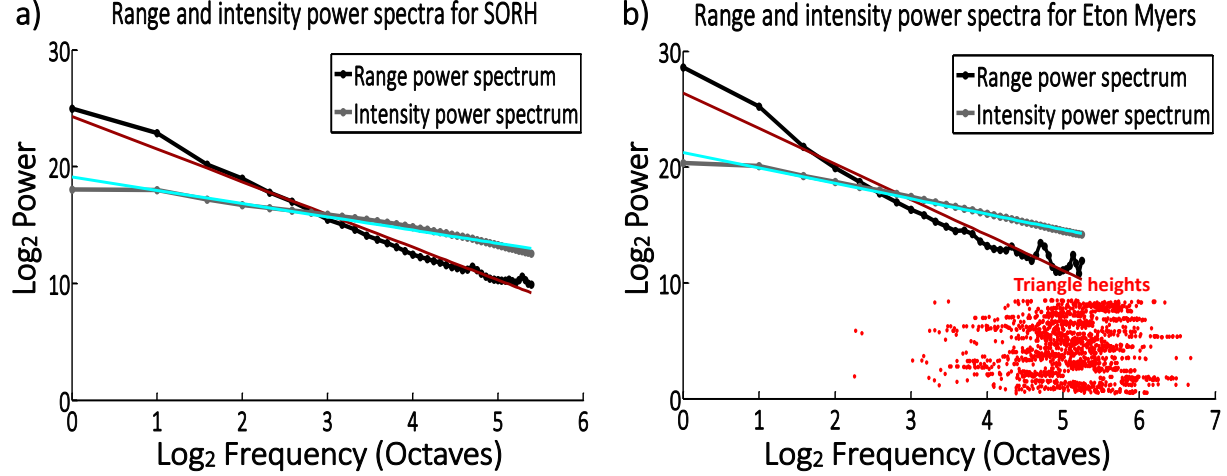


Figure 4.2: **a)** Plot showing the average power spectrum along with the power law fits for the SORH color and range images. **b)** Plot showing the average power spectrum along with the power law fits for the Eton Myers color and range images. The red dots mark the spatial frequencies whose wavelengths correspond to the average height of the triangles that comprise the 3D mesh of the object face which when projected onto image space makes the range image. There are 1604 dots corresponding to the 1604 range image renderings of the Eton Myers objects. The triangular facets of the SORH objects correspond to frequencies beyond the limit of the graph.

the vector or pixels prior to the power spectrum computation. All such power spectra are averaged together to provide $H_{cnts}(f)$.

It is important not to measure the power spectrum beyond the available resolution of the 3D objects. For each database, 3D shape is initially represented as a triangular mesh; we must verify that these triangular faces are small in height (number of pixels) in the rendered range images. For the SORH database, the triangular faces were on average 0.81 pixels in height. For the Eton Myers images, the average triangle height was 2.1 pixels.

Figure 4.2 shows the log-log plots of the power spectrum for the SORH and Eton Myers databases. For the SORH database, the 1D periodogram of intensity images had a slope of -1.14 , while the periodogram of range images had a slope of -2.8 , close to the predicted values of -1 and -3 respectively. For the Eton Myers database, the periodogram of in-

tensity images had a slope of -1.33 , while the periodogram of range images had a slope of -3.06 . We also found 95% confidence intervals for each fit: the slopes for the intensity and range periodograms for the SORH database have confidence intervals of $[-1.21, -1.07]$ and $[-2.93, -2.61]$, respectively, and $[-1.38, -1.28]$ and $[-3.31, -2.82]$ for the Eton Myers database.

In figure 4.2, we also show the frequencies whose wavelengths correspond to the average sizes of triangles comprising the 3D objects. For the SORH dataset, these frequencies all appear beyond the right edge of the figure.

We use simple linear regression of the log power spectrum to estimate the power law exponent α in $1/f^\alpha$. Although linear regression is commonly used to estimate the parameter and quality of fit of a power law for power spectra, this method can be sensitive to noise in circumstances with smaller sample sizes [95], and MLE methods are preferred in those cases. For larger sample sizes such as those used here, MLE estimates should agree with linear regression. In [96], the author proposes a technique which is a simple correction to the linear regression estimate of the power law exponent whose empirical performance is close to the maximum likelihood estimator. Using this correction technique, we verified that the linear regression estimates are accurate within two decimal places for all measurements.

The results observed are in agreement with the hypothesis that the range images of objects under roughly orthographic projection and viewed without self-occlusions should produce a 2D power spectrum that obeys a $1/f^4$ power law, while intensity images within unoccluded regions have a power spectrum that is similar to that of cluttered scenes.

4.5 Study of scaling properties of 3D objects

In section 4.4, studying the statistics of range images of unoccluded surfaces within single objects revealed that objects themselves exhibit fractal self-similar behavior in their 3D geometrical structure. Self-similarity of the shapes of objects has been hypothesized in the

past, but has not been directly observed within arbitrary natural objects. In particular, this observation can't be concluded solely from the scale invariance of images or range images of cluttered natural scenes, since occlusion alone is sufficient to explain scale invariance in cluttered scenes. Note, though, that the scale invariance found in the surfaces of objects may not necessarily imply that the full 3D shape of single objects shows scale invariance. The hypotheses of section 4.3.2 resulted from considering the consequences if objects exist within cluttered scenes that are scale invariant. Because individual objects have finite size, individual objects cannot exhibit scale invariance in low frequencies that fall below wavelengths that resemble the object's height. Still, an object may exhibit scale invariant geometry within it's higher frequencies; trees and ferns are commonly used examples. However, an object may have surfaces that show scale invariance without showing scale invariance over it's entire geometry. For example, the surface of the Earth shows self-similar structure [97], and yet the shape of the planet is well approximated by a spheroid, and does not scale.

The availability of full 3D models in the Eton Myers database allows us to investigate the scaling properties of whole 3D objects. The 3D models are in the form of object files which contain the set of vertices and triangles that make up the external surface of the object. We characterize each object using a binary volumetric representation $b(x, y, z)$ that represents object inclusion: $b(x, y, z) = 1$ for all points (x, y, z) contained within the object, and is zero elsewhere. In this section we analyze the scaling properties of these binary volumetric representations.

First, we established a single bounding box capable of enclosing all objects from the Eton Myers data set, plus a threefold margin to ensure that each object was entirely contained within this box. For each object, lines were cast across the object from the left edge of the object bounding box to the right edge. Each line intersects the object surface an even number of times (tangential intersections are discarded). For each line we located each intersection between the line and the surface; the points that fall in between these intersection points are classified as within the object. This allows us to compute $b(x, y_i, z_i)$ for fixed values y_i and z_i .

We then take the 1D power spectrum of $b(x, y_i, z_i)$ (after applying a Hanning window), and compute a cumulative power spectrum by averaging across the spectrum of each $b(x, y_i, z_i)$ for a densely sampled array of points y_i and z_i . The density of the line sampling was chosen so that each triangular facet that comprised the 3D mesh of the objects contained an average of 12 cast rays. This ensures that even a small facet is hit by at least one ray.

By equation 4.13, if objects are fully scale invariant ($\eta = 0$), then we expect the power spectrum of 1D slices through b to have a power spectrum of $1/f$. In contrast, if the power spectrum of each object is dominated by its coarse overall shape, we may expect its power spectrum to resemble that of a smooth sphere. A sphere is not scale invariant, and its power spectrum does not obey a power-law. We may expect the result to be close to the power spectrum of a step function, which is $\text{sinc}(f)^2 = \sin(f)^2/f^2$.

Figure 4.3c shows the power spectra of all Eton Myers objects. For most objects, the power spectrum oscillates substantially and therefore is not well fit by a power-law. Overall, the rate of drop-off in power is close to $1/f^2$, closely resembling the power spectrum of a sphere. This suggests that the individual man-made artifacts of the Eton Myers dataset are not fully scale invariant in shape over any significant frequency band. Thus, full 3D structural scale invariance is not necessary for the range images of objects to obey 3D scale invariance within occlusion contours.

We can compare these power spectra to that of an artificial object that is generated to achieve full scale invariance. The object in figure 4.3a was constructed using a 3D variation of the collage model. Recall that the collage model generates scale invariant 2D images by iteratively overlaying 2D objects on top of the scene, at random locations, random sizes, and random color. The resulting image is scale invariant if and only if the area of the 2D objects obey a $1/r^3$ Pareto distribution [79]. Using a similar method, a 3D scale invariant object can be constructed by initializing voxels in 3D cubical space to an unlabeled state, and then in each iteration randomly placing a sphere within the space and labeling all voxels within the sphere as either one (occupied space) or zero (unoccupied space). This process repeats until

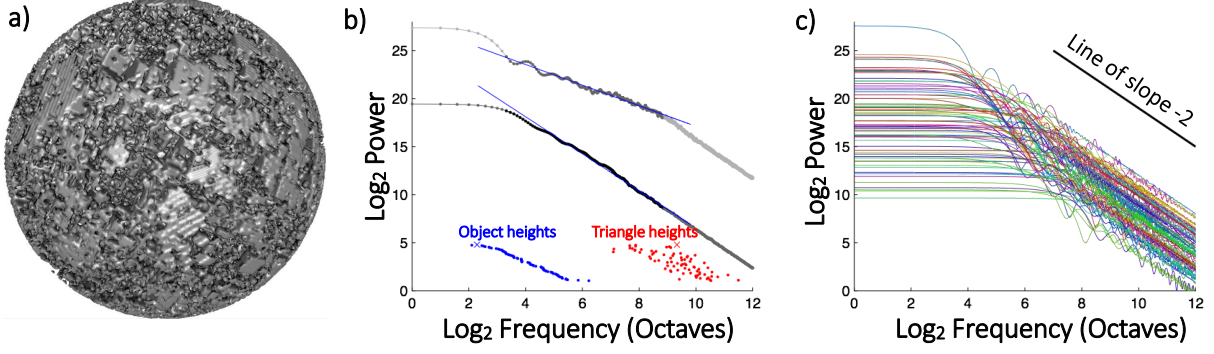


Figure 4.3: **a)** One example of a fully scale invariant object. While there are many ways to construct such an object, this particular example is generated using a 3D dead leaves model. The object does not have a solid interior; it's complex structure of holes and crevices extends throughout its mass, like a sponge. **b)** Plot showing average power spectra of Eton Myers objects (black), compared against the power spectra of an artificial scale-invariant object (shown in gray). Along the bottom, blue dots mark the spatial frequencies whose wavelengths correspond to the heights of the objects, and red dots mark the frequencies that correspond to the average width of the triangles that comprise the 3D mesh representing each object. The vertical positions of these marks don't carry meaning; they are arranged vertically for clarity only, sorted according to the object size for illustration purposes. The blue and red crosses show the width and triangle size for the artificial scale-invariant object. **c)** The power spectra of 1D slices of $b(x, y, z)$ for all 74 Eton Myers objects. These spectra do not obey a power law, and their power decreases at a rate similar to $1/f^2$.

all voxels are labeled. Once the process has filled the bounding box, we truncate the shape to a sphere to ensure that the object has a finite size similar to objects in the database.

The resulting shape will be fully scale invariant ($\eta=0$) if the radii of the randomly-placed spheres are drawn from a $1/r^4$ Pareto distribution. One intuitive way to derive this is by constraining the 3D Poisson point process so that the expected number of spheres between one and two units in height that lie within a neighborhood of a fixed number of units in diameter should not depend on the choice of units. In other words, if an observer within this universe was reduced in size, the expected number of spheres of similar size to the observer that lie within a few steps in any direction should remain constant as the observer shrinks. The expected number of objects within a spherical neighborhood of diameter r is proportional to r^3 , and the percent of objects that lie within heights Ar and Br for some constants A and B is given by $\int_{Ar}^{Br} p(r)dr$. Thus, if the 3D collage is scale invariant,

then $\int_{Ar}^{Br} p(r)dr \propto 1/r^3$. By differentiating both sides and letting $A \rightarrow \infty$, we have that $p(r) = 1/r^4$. Note that under pinhole camera geometry, if the heights of the 3D spheres was drawn from $1/r^4$, then the heights of the 2D images cast by the random spheres would obey a $1/r^3$ distribution, which matches the constraint required for 2D collage images to be fully scale invariant.

The power spectrum for this object is shown in figure 4.3b (in light gray), and is measured as having a slope of -1.07 . Here, we measure the power spectrum within a frequency range bounded in the low frequencies by wavelengths that are within an octave of the height of the object, and bounded in the high frequencies by wavelengths that are within an octave of the size of the triangles that comprise the object's 3D mesh. This range is darkened in the graph of figure 4.3b.

In figure 4.3b, we also plot the average power spectrum of all objects. This spectrum closely follows a $1/f^{-1.93}$ power law ($\eta \approx -1$). Given that the individual spectra dropped off in frequency as quickly as $1/f^2$, the mean power spectrum would only obey $1/f$ scale invariance if the object heights were distributed as $p(r) = 1/r^4$ (or, equivalently, if object volumes were distributed as $p(v) = 1/v^2$). The distribution of sizes of objects in the Eton Myers data set was not easily described by a power law; a larger sample would be required to estimate the distribution of sizes of man-made objects.

4.6 Discussion

In section 4.4, we find that the range images of unoccluded regions of the surfaces of single objects show $\eta = -2$ self-similarity in their covariance structure, which suggests that the geometrical shape of single objects tend to exhibit self-similarity along their surfaces. This self-similarity is consistent with 3D scale invariance of cluttered scenes, as shown in section 4.3.2. We further show in section 4.5 that only the surfaces of objects show self-similarity; the 3D power spectra of volumetric representations of typical single objects appear to be

dominated by their finite size and coarse overall shape. This suggests that single objects may be better modeled as a simple 3D shape overlaid with self-similar texture (such as a topographical relief globe of the Earth) rather than a volumetrically self-similar shape such as a fern, or a self-similar sponge like the 3D collage shape in figure 4.3a.

These findings were consistent between the SORH database (which included a mixture of small manmade and natural objects) and the Eton Myers database (which included only manmade museum artifacts). Still, variations in the statistics of objects surfaces can be expected for selections of objects that are significantly different from these (for example, plants and trees may exhibit different statistical structure).

These findings suggest that range images of cluttered scenes may exhibit self-similarity at two different scales simultaneously, with a $\eta=0$ scale invariance caused by occlusion that overshadows a self-similarity of scale $\eta = -2$ that occurs with a much smaller amplitude along the surfaces of objects.

We also show in section 4.4 that unlike range images, the power spectrum of the *appearance* of the unoccluded surfaces of single objects shows $\eta=0$ scale invariance. It has been extensively debated in the literature whether the $1/f^2$ power spectrum of cluttered images is due entirely to occlusion, if it is merely the consequence of the prominence of sharp lines, or some combination thereof [87, 88, 98]. The results here show that, although the effects of occlusion may well dominate the statistics of cluttered scenes, the $1/f^2$ power spectrum of natural images continues even in the absence of any occlusion. Note, however, that section 4.3.2 predicts that unoccluded surfaces with little variation in albedo, no cast shadows, and oblique point-source lighting should produce a power spectrum that is $1/f^2$ along the dominant lighting direction, but $1/f^4$ along the direction perpendicular to the lighting. More research would be required to determine if such shading effects are significant in natural scenes.

The observation that the range images of single unoccluded surfaces show $\eta=-2$ power spectra has immediate applications in computer vision. The inference of 3D shape is often

highly ambiguous. Classical computer vision problems of binocular and multi-view stereo, photometric stereo, shape from shading, structure from motion, and others all rely on accurate and complete models of which 3D shapes and textures are plausible in nature, to avoid producing unlikely outputs. Bayesian approaches are common for these problems, and depth inference problems are often decomposed according to Bayes rule:

$$P(shape|images) \propto P(images|shape)P(shape) \quad (4.26)$$

First, consider those applications that seek to infer the shape of a single surface. Applications for single object depth inference include robotic grasping, astronomy, medicine, and computational photography. While earlier methods of probabilistic inference were constrained to pairwise connected Markov random fields in order to remain tractable, more recent advances enable efficient inference using fully connected spatial priors [41,99–101]. Fully connected priors are capable of exploiting the entire covariance structure of the Bayesian prior. Among all spatial priors that exhibit a specific covariance structure, the prior with the maximum entropy (which can therefore be considered to make the fewest additional assumptions) is a Gaussian:

$$P(\vec{z}) \propto \exp -\vec{z}'C\vec{z} \quad (4.27)$$

where C is the desired covariance matrix. If the desired covariance structure is $1/f^\alpha$, then C can be expressed as $C = F'DF$, where F is a Fourier basis, and D is a diagonal matrix whose diagonal values are specified by $1/f^\alpha$. In [99], expectation propagation was used to solve shape from shading problems using a $1/f^2$ covariance structure, chosen to match the covariance structure observed for range images of cluttered scenes. The findings of section 4.4 suggest that a $1/f^4$ prior might be more appropriate for unoccluded surfaces. Expectation propagation is rarely applied to computer vision because it historically requires estimating the covariance matrix of the posterior distribution, which is an expensive operation for dense

pixel-level computer vision applications because the number of elements in the covariance matrix is quadratic with the number of pixels in the image. In [99], it was shown that assuming rough approximation of the covariance structure (here, a Gaussian with a $1/f^2$ power spectrum was used) was sufficient to avoid the expensive process of estimating the covariance structure, and thereby enable an otherwise intractable inference algorithm for dense pixel-level vision, providing state of the art performance on shape from shading. We hope that findings on the statistics of the shape of single objects from this work and others will both inform new and more accurate Bayesian priors on shape, and also enable more efficient probabilistic inference procedures.

For applications that infer surfaces with significant self-occlusion, $P(shape)$ may be modeled using an explicit occlusion process, where the locations of occlusion contours are represented as a binary image. In this circumstance, $P(shape)$ might impose a $1/f^4$ covariance structure over regions within the explicit occlusion contours. Finally, applications that infer the full 3D shape of an object might benefit from spatial priors that model single objects as a coarse 3D shape covered by a $1/f^4$ texture.

Chapter 5

Statistical Correlations Between Intensity and Range Image Pairs of Single Objects

5.1 Introduction

As mentioned in the earlier chapters classic depth inference algorithms rely on many simplifying assumptions which are unrealistic and hence do not generalize well towards natural scenes. The study of statistical relationships between intensity and range images for depth inference is slowly gaining importance as seen in [4, 19, 21]. The papers on "make3D and automatic photo pop-up" [28, 30, 31, 43] focus mainly on outdoor scenes, with the camera oriented parallel to the ground. The statistics of such scenes are highly regular due to the ground plane and the tendency for any structures built on the ground to be vertically oriented. Depth inference algorithms in such environments can rely heavily on those regularities. Potetz and Lee [4] collected statistics from natural imagery consisting of both urban and rural scenes and reported a striking correlation between nearness and brightness which they call "da Vinci correlation". But this correlation depends on the presence of complex

3D surfaces with crevices and concavities, cast shadows, diffuse lighting and lighting inter-reflections. In [19], Torralba and Freeman proposed "Shape Recipes" which are nothing but linear regression kernels capturing the relationship between intensity and range image pairs. Potetz and Lee in [21] extend this concept and state that the linear relationship between intensity and range image pairs for outdoor natural images could be approximated given three parameters: the strength of the shading cue for that scene, the strength of the shadow cue (nearness/brightness correlation) for that scene and the main lighting direction. The focus of both these works was to infer high resolution 3D shape given a high resolution intensity image and only a low resolution range image. In this chapter, we generalize this approach to the case of depth inference (without relying on any low resolution range image) of single objects and surfaces photographed under controlled illumination. Again the three parameters might not hold the same significance in this case, like that for natural scenes. For these reasons we investigate into the statistics of images of single objects and surfaces and their range data from the SORH and Eton Myers databasets which might throw some new insights.

5.2 Mathematical Background

5.2.1 Linear Regression Model

The mathematical model in this section is reproduced from [21] by Potetz and Lee. We start with the cross-correlation between the image $i(x, y)$ and range image $z(x, y)$ as follows:

$$(i * z)(\Delta x, \Delta y) = \iint i(x, y) z(x + \Delta x, y + \Delta y) dx dy \quad (5.1)$$

Let $I(u, v)$ and $Z(u, v)$ is used to denote $i(x, y)$ and $z(x, y)$ in the Fourier domain, then the Fourier transform of $i * z$ is $Z(u, v)I^*(u, v)$ which is known as the cross spectrum of i and z . Given this cross spectrum for an image which is sufficiently bounded away from zero, the

3D shape can be estimated from a single image using linear regression as follows:

$$Z = I(ZI^*/II^*) \quad (5.2)$$

If we can estimate the regression kernel given by $K = ZI^*/II^*$, the shape can be perfectly reconstructed using equation 5.2. It is observed that the cross spectrum ZI^* for outdoor natural images can be fit by a power law modeled by $B(\theta)/r^\alpha$ where r is the spatial frequency in polar coordinates, $B(\theta)$ is a parameter of the model which depends on the polar angle θ and α was measured to be approximately 1 from their dataset.

5.2.2 Pentland's Linear Lambertian Model

Starting with the Lambertian model of image formation of a surface $z = z(x, y)$ given as:

$$i(x, y) = R(p, q, \mathbf{L}) \quad (5.3)$$

where $i(x, y)$ is the corresponding image formed, R is the reflectance function which depends on the slopes of the surface p, q and the illumination direction \mathbf{L} given simplifying assumptions such as uniform albedo, orthographic projection and no shadowing.

Pentland [32] introduced a technique which forms a linear approximation of the reflectance function in terms of p and q . This is done by taking a Taylor expansion of R around $(p, q) = (p_0, q_0)$:

$$R(p, q) = R(p_0, q_0) + (p - p_0) \left. \frac{\partial R(p, q)}{\partial p} \right|_{p=p_0, q=q_0} + (q - q_0) \left. \frac{\partial R(p, q)}{\partial q} \right|_{p=p_0, q=q_0} \quad (5.4)$$

Pentland showed that this linear approximation is quite accurate over a sufficiently small range of p and q values and this range increases as the illuminant direction is more oblique with respect to the viewer.

For lambertian reflectance and $(p_0, q_0) = (0, 0)$ Eq. 5.4 becomes:

$$i(x, y) = \rho\lambda[\cos\sigma + p\cos\tau\sin\sigma + q\sin\tau\sin\sigma] \quad (5.5)$$

where ρ is the albedo of the surface, λ is the strength of the illuminant at the surface, τ is the tilt of the illuminant and σ is the slant.

Eq. 5.5 can be written as

$$i(x, y) = k_1 + pk_2 + qk_3 \quad (5.6)$$

where $k_1 = \cos\sigma$, $k_2 = \cos\tau\cos\sigma$ and $k_3 = \sin\tau\sin\sigma$ and so $L = (k_2, k_3, k_1)$ form the generalized illuminant direction.

Taking Fourier transform of Eq. 5.5 on both sides we get

$$I(r, \theta) = 2\pi r j[k_2 \cos\theta + k_3 \sin\theta]Z(r, \theta) \quad (5.7)$$

where $j = \sqrt{-1}$ and hence we have:

$$ZI^*(r, \theta) = \frac{1}{2\pi r j[k_2 \cos\theta + k_3 \sin\theta]}II^*(r, \theta) \quad (5.8)$$

$$K(r, \theta) = -j \frac{1}{r} \frac{1}{2\pi[k_2 \cos\theta + k_3 \sin\theta]} \quad (5.9)$$

The $1/r$ drop-off in the imaginary part of K in section 5.2.1 is attributed to Pentland's linear Lambertian model of shading [32] shown in section 5.2.2.

Hence Pentland's model also predicts that $\text{imag}[ZI^*]$ should obey a $1/r$ power-law as observed from the statistics of the natural image and range database of Potetz and Lee. But according to this model, the real part of ZI^* should be zero which is contradictory to the observations made. Studies conducted on their natural image database show that the real part is actually stronger than the imaginary part. This is explained by the fact that

the real part is the Fourier transform of the even symmetric part of the cross-correlation function and corresponds to the direct correlation between nearness and brightness and hence is stronger than the imaginary part. Potetz and Lee in [4] report significant negative correlations between intensity and range images of natural scenes; nearby pixels appear brighter and vice versa. This correlation which they called the "da Vinci correlation", is attributed to the complex cast shadows in cluttered scenes containing many folds and concavities like foliage. The real part of $B(\theta)$ was shown to be consistently negative which points in the same direction.

While Pentland's linear Lambertian model predicts that the real part of ZI offers no useful correlations between intensity and range images, Potetz and Lee in [21] show that for cluttered natural scenes comprising of concavities and complex cast shadows, the real part was stronger. The main goal in [21] was to construct high resolution range images given only low resolution range and high resolution intensity images. They show that a significant part of the improvement achieved in depth inference was due to shadow cues. When the real part of the linear regression kernel K was suppressed, their algorithm achieved far less improvement than when the real part was also used. While all this holds for natural scenes, the question is whether they are applicable to images of single objects and surfaces and their range data. The following sections investigate into the correlational statistics of intensity and range images of single objects from the SORH and Eton Myers datasets.

5.3 Experimental Methodology

We start our analysis by investigating the statistical correlations that exist between intensity and range images of single objects using linear ridge regression. The intensity and range image pairs from the SORH and Eton Myers datasets were used for this purpose. Color images were converted into gray-scale intensity images by averaging over the three color channels. Experiments were conducted using two approaches:

1. the patch based method and
2. the pair-wise method

5.3.1 The patch based approach

For this study, 25×25 patches of coregistered intensity and range data were used. There are a total of 1,197,932 patches from the SORH dataset and 308,925,107 patches from the Eton Myers dataset. Only pixels falling on/within the object boundaries were considered valid. Any patches containing invalid pixels (i.e., background pixels and pixels for which range data is unavailable) were discarded. The main drawback of this approach is the loss of patches falling on object boundaries since part of them are invalid pixels.

Let i is a 25×25 intensity patch and z be the corresponding range patch. Let N be the number of patches in our dataset. Then:

$$\bar{i}(x, y) = \frac{1}{N} \sum_{k=1}^N i_k(x, y) \quad (5.10)$$

$$\bar{z}(x, y) = \frac{1}{N} \sum_{k=1}^N z_k(x, y) \quad (5.11)$$

where x, y represents the (x, y) th pixel in the patch, \bar{i} and \bar{z} are the patch averages and i_k and z_k are the k th patch in the patch dataset. The average patch is subtracted from each patch in the dataset:

$$\hat{i}_k = i_k - \bar{i} \quad (5.12)$$

$$\hat{z}_k = z_k - \bar{z} \quad (5.13)$$

The mean subtracted patch matrices are then vectorized and stored in intensity and range matrices. Let \mathbf{I} and \mathbf{Z} be the $N \times 25^2$ matrices containing all the mean subtracted vectorized intensity and range patches where each row corresponds to a single patch. Then $\mathbf{I}'\mathbf{I}$ and $\mathbf{Z}'\mathbf{Z}$ are the 625×625 auto covariance matrices of the intensity and range patches. Similarly $\mathbf{I}'\mathbf{Z}$ is the cross covariance matrix. The Pearson's correlation coefficient is then computed as:

$$\rho = \frac{cov[\mathbf{I}, \mathbf{Z}]}{var[\mathbf{I}]var[\mathbf{Z}]} \quad (5.14)$$

Fig. 5.1 shows a color coded 2D rendering of the correlation between intensity at pixel (13, 13) (which is essentially the middle pixel in the patch) and all the pixels of the range patch for all lighting directions.

5.3.2 The pair-wise approach

In this approach, statistics are computed for pairs of individual pixels falling within a fixed size window. This method has the advantage that boundary pixels are not discarded as in the patch based approach. Let $i_k(x, y)$ and $z_k(x, y)$ are the intensity and range values at pixel location (x, y) in the k th image in the dataset. Imagine a 201×201 window centered around this pixel. Let $i_k(l, m)$ and $z_k(l, m)$ are the intensity and range values at pixel location (l, m) in the same image falling within the window. We compute a number of quantities related to the covariance structure for this pixel pair at (x, y) and (l, m) and store it in a 201×201 matrix \mathbf{M} at the location $(l - x, m - y)$. Statistics are computed for all possible pixel pairs within the window and added to the corresponding location in the \mathbf{M} matrix. The window based process is repeated at every valid pixel location in all the images and the results added

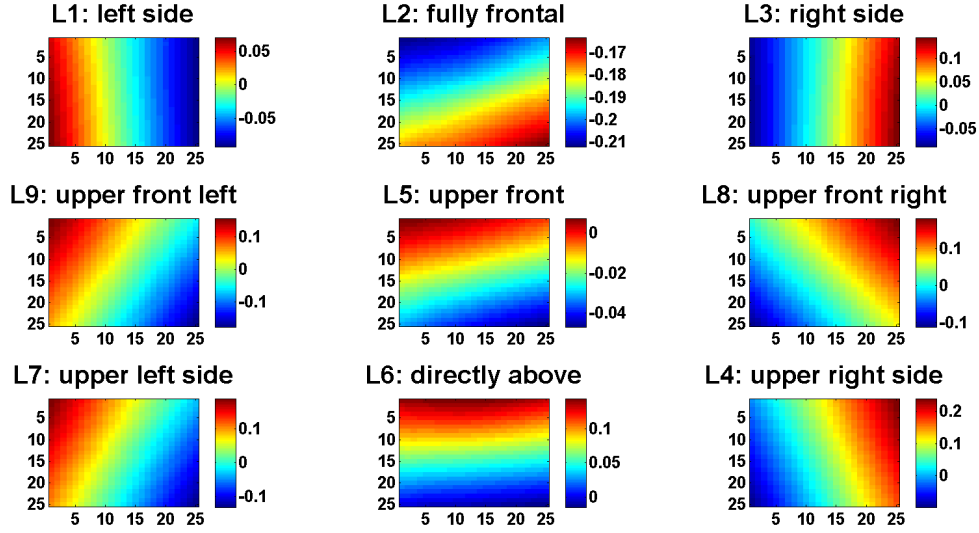


Figure 5.1: Color coded 2D rendering of the correlation between intensity at pixel (13, 13) and all the pixels of the range patch for the 9 light source directions for the SORH dataset. The light sources are named according to their placement with respect to the object. The first row corresponds to the light sources which are in the same horizontal plane as the object. The second row corresponds to the light sources which are in front of the object but the light source-object vector is oblique with respect to the camera-object plane. The third row corresponds to the light sources which are in the same vertical plane as the object. Refer to fig. 3.2 for a graphical illustration of the light sources (L1 - L9), camera and object setup.

to the corresponding locations in \mathbf{M} . Each location (p, q) in the matrix \mathbf{M} represents the aggregate statistics of pixel pairs separated by the distance p in the horizontal and distance q in the vertical directions across all images in the dataset. The following statistics are recorded:

- $\mathbf{M}_{\mathbf{I}_1 \mathbf{I}_2}$ - covariance between pixel pair intensities where \mathbf{I}_1 and \mathbf{I}_2 represent the intensity values of the pixel pairs,
- $\mathbf{M}_{\mathbf{Z}_1' \mathbf{Z}_2}$ - covariance between pixel pair range values where \mathbf{Z}_1 and \mathbf{Z}_2 represent the range values of the pixel pairs,
- $\mathbf{M}_{\mathbf{I}_1 \mathbf{Z}_2}$ - covariance between intensity pixel and the range pixel from its corresponding pair and

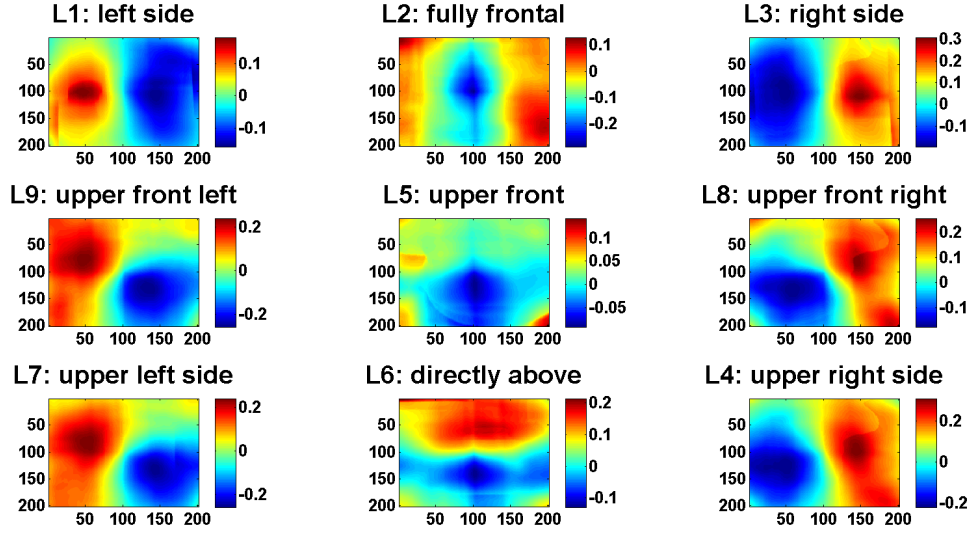


Figure 5.2: Color coded 2D rendering of the correlation between intensity and range pixels separated by distances ranging from 0 to 100 in either direction for the 9 light source directions for SORH dataset

- $\mathbf{M}_{I_1 Z_1}$ - covariance between intensity and range pixels at the same location.

Similarly correlational statistics are computed and Figs. 5.2 and 5.3 show the color coded 2D renderings of the correlations for the SORH and Eton Myers datasets respectively.

Potetz and Lee [4], in their study of statistics between intensity images of natural scenes (containing both rural and urban scenes) made a striking observation that shadow cues constitute a significant part of the observed statistics which are attributed to the presence of cluttered scenes (like foliage) containing deep crevices and concavities which tend to be in shadow. These manifested in the form of significant negative correlations suggesting that nearby pixels tend to be brighter and vice versa. In the SORH and Eton Myers datasets of single objects, no significant negative correlations were observed. The correlations observed ranged between 0.3 and -0.2. We see relatively high correlations in the direction of incident light which fades as we move away from the light source as illustrated in figures 5.1 and 5.2. The images in both our datasets contain no deep folds or crevices which might contribute to significant negative correlations. The conclusion based on the evidence collected from the

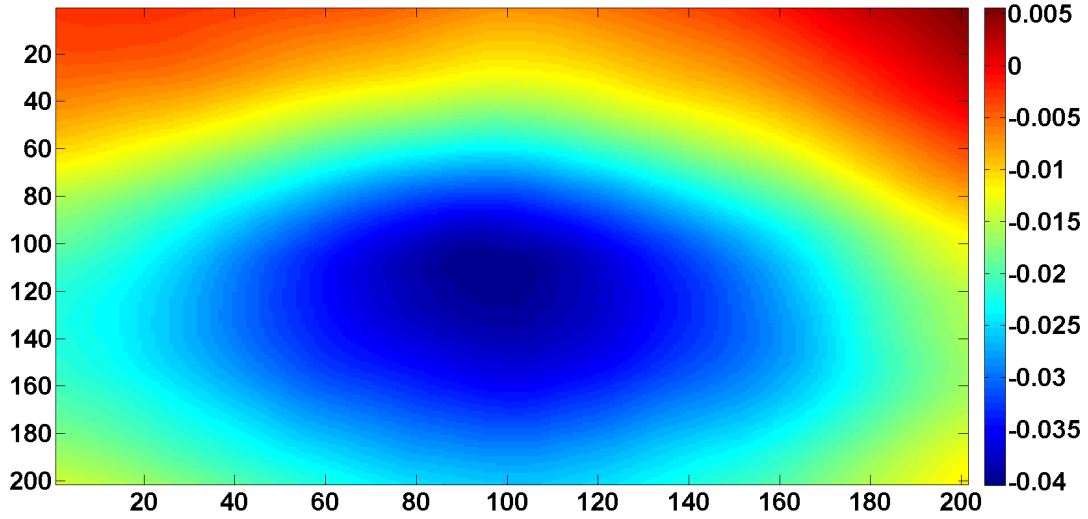


Figure 5.3: Color coded 2D rendering of the correlation between intensity and range pixels separated by distances ranging from 0 to 100 in either direction for the Eton Myers dataset

two datasets corroborates the finding that the nearness/brightness correlation observed in cluttered scenes stems from cast shadows. Hence, $B(\theta)$ depends mainly on shading cues and the dominant illuminant direction for scenes containing single objects.

In chapter 4, it is shown both analytically and experimentally that the range images of single objects without occlusions has a $1/f^3$ power law 1D power spectrum. The same is tested here using the cross correlation matrices. The power spectrum computed by taking the 2D Fourier transform of the cross-correlation matrix $\mathbf{M}_{\mathbf{z}_1' \mathbf{z}_2}$, is fit by a power law having slope -2.8 for the SORH dataset and -3.16 for the Eton Myers dataset. The slopes are an average of power law fits computed over horizontal and vertical slices taken from the middle row and column of the power spectrum matrix. These are in agreement with power law fits computed over the power spectra of range images for the SORH and Eton Myers datasets as seen in chapter 4 fig. 4.2.

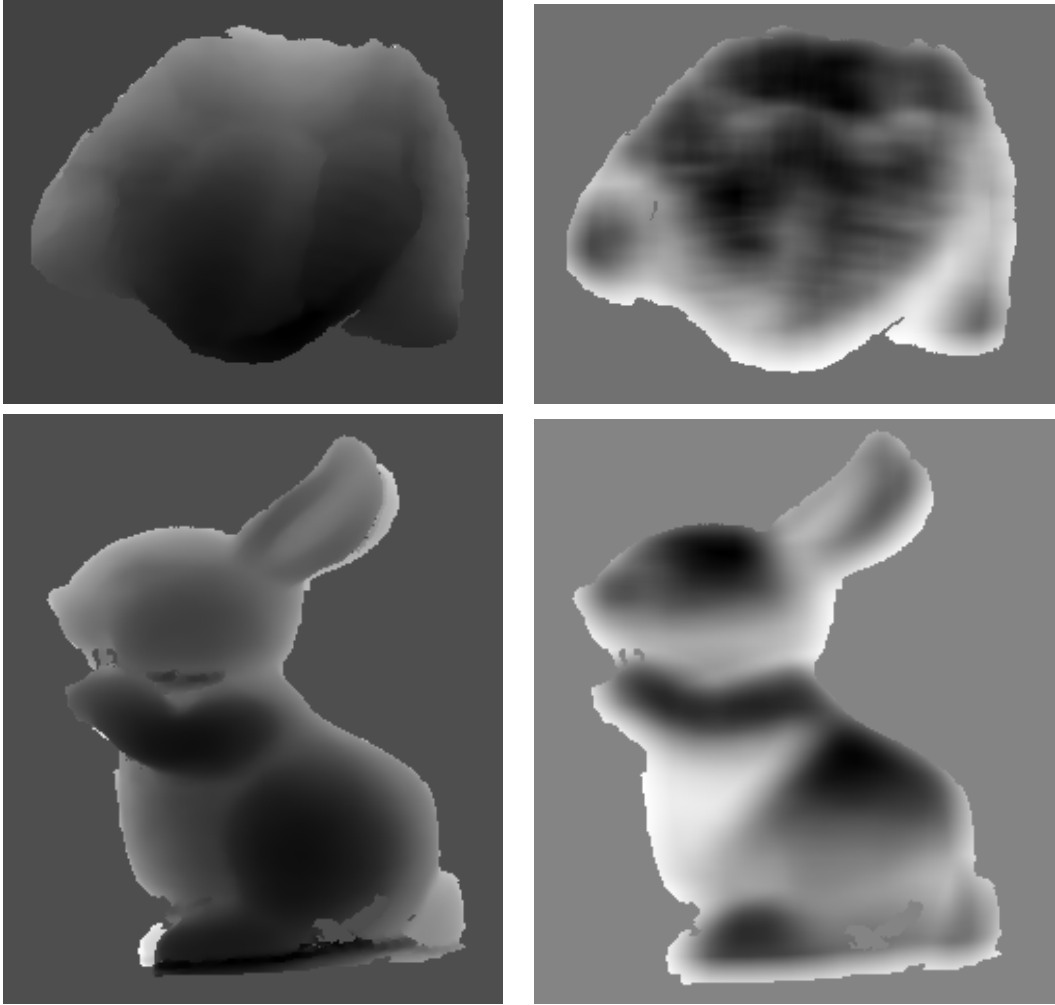


Figure 5.4: Range images of ground truth (first column) and predicted shape (second column) using linear regression of two objects from the SORH database.

5.4 Depth Inference Using Ridge Regression

As a first attempt, we used a simple ridge regression algorithm for depth inference using the cross covariances computed between intensity and range values using the patch based approach in section 5.3.1

$$\mathbf{K} = (\mathbf{I}'\mathbf{I} + \lambda\mathbf{1})^{-1}\mathbf{I}'\mathbf{Z} \quad (5.15)$$

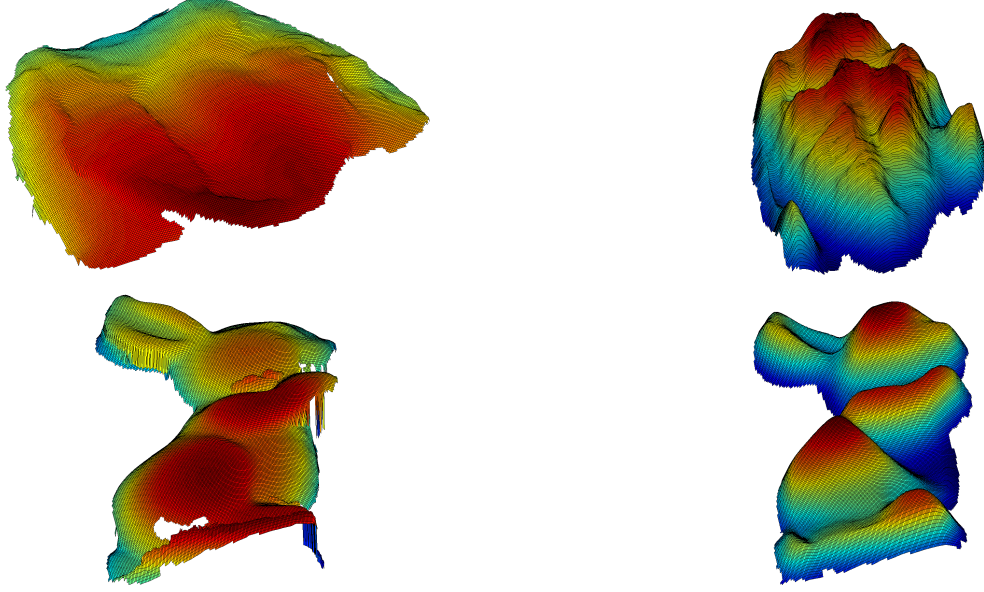


Figure 5.5: 3D mesh plots of ground truth (first column) and predicted shape (second column) using linear regression of two objects from the SORH database.

where $\mathbf{I}\mathbf{I}$ and $\mathbf{I}\mathbf{Z}$ are the 625×625 auto covariance and cross covariance matrices respectively, $\mathbf{1}$ is an identity matrix and λ is a regularization parameter. \mathbf{K} is a 625×625 parameter matrix learned from the data. A single column was extracted from the middle of the \mathbf{K} matrix and applied as a 25×25 patch filter \tilde{K} to the intensity images to predict the corresponding range images:

$$\hat{z} = \tilde{K} * i \quad (5.16)$$

Fig 5.4 shows the ground truth and predicted depths for two objects from the SORH dataset. Fig. 5.5 compares the 3D mesh plots of the same. The estimated shapes look reasonable given the fact that the training set consists of a variety of objects of different albedos and surface reflectance properties. Also the linear regression kernel is heavily influenced by the dominant lighting direction which is evident from the results. Any possible improvement in the estimated shape might need a much larger training set with more examples for each category of albedo and surface reflectances.

The ridge regression based depth inference algorithm is used as a baseline to compare the Expectation Propagation based shape inference algorithm in Chapter 6. The statistical insights gained about single objects and surfaces in chapter 4, suggests a prior that exhibits a $1/f^4$ covariance structure.

Chapter 6

SFS Using Expectation Propagation

6.1 Probabilistic Graphical Models in Vision

Probabilistic inference problems arise in many scientific fields like computer vision, artificial intelligence and physics. Many problems in computer vision can be modeled as large pairwise connected Markov Random Fields and there has been continued effort to develop better inference algorithms. Belief Propagation (BP) is one such algorithm which has been successfully applied to many computer vision problems [41, 102–104]. The main drawback of belief propagation is that its complexity grows exponentially with the clique size due to which interactions are limited to pairwise, as anything beyond that gets incredibly expensive. In 2001, Minka proposed the Expectation Propagation (EP) algorithm [105]. EP is a message passing algorithm and is a generalization of BP. EP approximates the posterior distribution as a multivariate Gaussian which substantially simplifies the calculations required for inference. Section 6.3 provides a brief overview of the EP algorithm.

6.2 Overview of Markov Random Fields

Probabilistic Graphical Models (PGM's) use a graph to represent a complex probability distribution over a high-dimensional space. The nodes in the graph represent the variables of

the problem while the edges represent the probabilistic dependencies between them. There are two types of PGM's: Bayesian networks or Directed Graphical Models and Markov Networks or Undirected Graphical Models. They differ in the kind of independencies that they can model and the kind of factorizations that they can induce.

Let $G = (V, E)$ denote a graph consisting of a set of nodes V and edges E . For each node $i (i \in V)$, let X_i denote the associated random variable. Let $\mathbf{X} = (\mathbf{X}_i)_{i \in V}$ denote the joint random variable. The nodes in such a graph form a Markov Random Field (MRF) if they satisfy the local Markov property:

1. **Pairwise Markov Property:** Any two non-adjacent variables are conditionally independent given all other variables:

$$X_u \perp\!\!\!\perp X_v \mid \mathbf{X}_{V \setminus \{u, v\}} \quad \text{if } \{u, v\} \notin E$$

2. **Local Markov Property:** A variable is conditionally independent of all other variables given its neighbors:

$$X_v \perp\!\!\!\perp \mathbf{X}_{V \setminus \text{cl}(v)} \mid \mathbf{X}_{\text{ne}(v)}$$

where $\text{ne}(v)$ is the set of neighbors of v , and $\text{cl}(v) = \{v\} \cup \text{ne}(v)$ is the closed neighborhood of v .

3. **Global Markov Property:** Any two subsets of variables are conditionally independent given a separating subset:

$$\mathbf{X}_A \perp\!\!\!\perp \mathbf{X}_B \mid \mathbf{X}_S$$

where every path from a node in A to a node in B passes through S.

According to Hammersley-Clifford theorem, for an MRF, the joint probability distribution can be factorized over the cliques of G as follows:

$$P(\mathbf{X} = \mathbf{x}) = \prod_{\mathbf{C} \in \text{cl}(\mathbf{G})} \phi_{\mathbf{C}}(\mathbf{x}_{\mathbf{C}}) \tag{6.1}$$

where $\text{cl}(G)$ is the set of maximal cliques of G and ϕ_C are the clique potentials.

A clique is a fully connected subset of nodes in the graph. A clique is maximal if it is not contained within any other larger clique. Clique potentials are functions over the nodes/variables spanning the clique.

For computer vision applications, MRF's can be used to model grid structured data (like natural images, depth maps, optical flow, etc). The interactions can be pair-wise or higher-order. In pair-wise MRF's, the clique potentials are defined on cliques of order strictly less than 3. They are popular in computer vision because of their simplicity and computational efficiency. However, pair-wise models are insufficient to encode the rich and complex statistical dependencies between variables. Hence, current research has been leaning towards exploiting higher-order interactions between variables which are defined on cliques containing more than two nodes/variables.

6.3 Overview of Expectation Propagation

Expectation Propagation (EP) [42, 105] works by approximating a factorized probability distribution

$$p(\vec{x}) = \frac{1}{Z} \prod_{i=1}^N \phi_i(\vec{x}_i) \quad \vec{x}_i \subseteq \vec{x} \quad (6.2)$$

by a simpler one:

$$\tilde{p}(\vec{x}; \vec{\theta}) \propto \exp\left(\sum_j \theta_j \psi_j(\vec{x})\right) \quad (6.3)$$

where $\tilde{p}(\vec{x}; \vec{\theta})$ belongs to an exponential family of distributions.

The parameter vector θ corresponds to a different weighting of the potential function

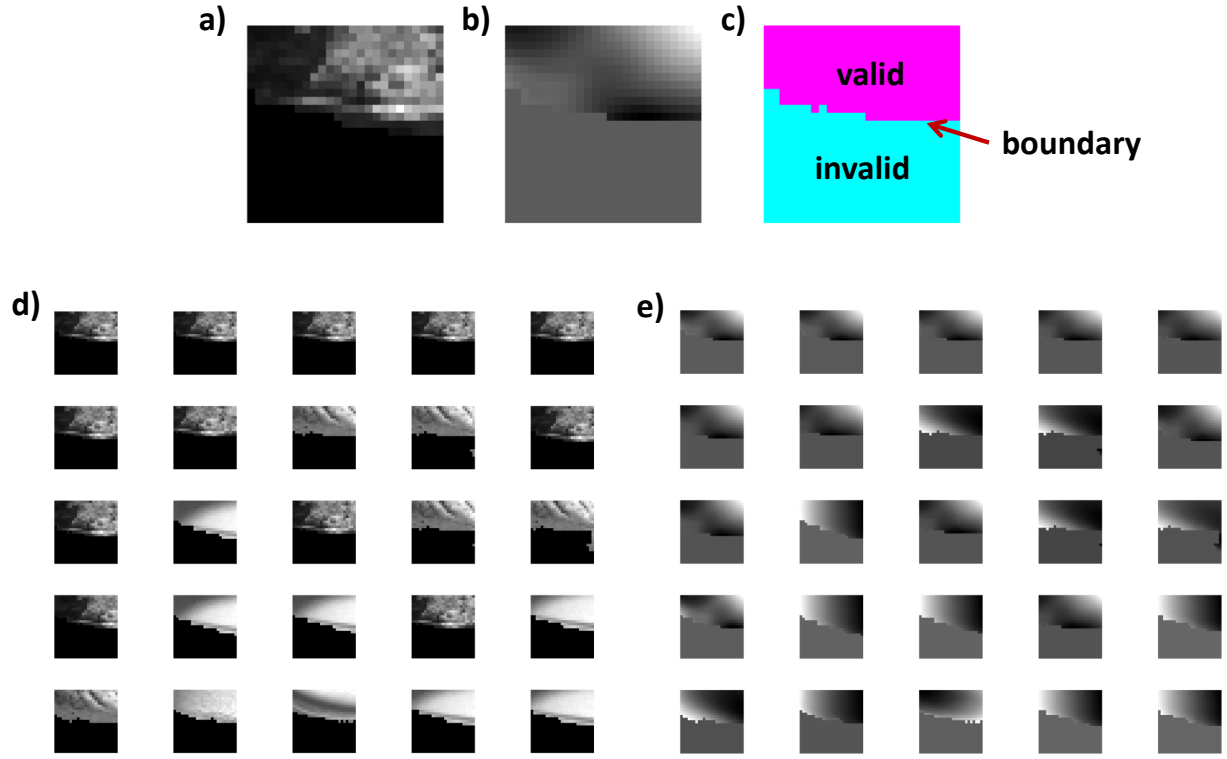


Figure 6.1: a) Query intensity patch, b) and c) corresponding query range and label patches. The valid and invalid pixels as well the boundary separating them are clearly marked in the label patch. d) Top 25 intensity patch matches along with their e) corresponding range patches.

$\psi_j(\vec{x})$.

EP finds the best approximation (\tilde{p}) to a distribution p by moment matching as follows:

$$proj[\phi_i(\vec{x}_i) \prod_{j \neq i}^N \tilde{p}_j(\vec{x}_j; \vec{\theta}^{(j)})] = \tilde{p}(\vec{x}; \vec{\theta}) \quad (6.4)$$

where $proj$ is a moment matching operator. EP reduces to BP if the approximating distribution is a product of discrete univariate marginals.

For exponential families, moment matching can be interpreted as minimizing the Kullback-Liebler (KL) divergence $D(P \parallel Q)$ where:

$$\begin{aligned}
P &= \phi_i(\vec{x}_i) \prod_{j \neq i}^N \tilde{p}_j(\vec{x}_j; \vec{\theta}^{(j)}) \\
&\text{and} \\
Q &= \tilde{p}(\vec{x}; \vec{\theta})
\end{aligned} \tag{6.5}$$

This process is repeated until the distribution hopefully converges which is not guaranteed but has been shown to work well in many applications.

In practice the approximating distribution is mostly chosen to be Gaussian with a distribution $\mathcal{N}(m_x, S_x I)$. We need to find the parameters m_x and S_x that best approximates $p(\vec{x})$ in eq. 6.2. For each factor i , iterate until convergence the following three steps:

1. Remove the factor ϕ_i from $p(\vec{x})$ and estimate the parameters of $p_{\setminus i}(\vec{x}) = \prod_{j \neq i}^N \phi_j(\vec{x}_j)$ as:

$$(S_x^{\setminus i})^{-1} = S_x^{-1} - S_i^{-1} \tag{6.6}$$

$$m_x^{\setminus i} = (S_x^{\setminus i})^{-1} (S_x^{-1} m_x - S_i^{-1} m_i) \tag{6.7}$$

2. Recompute the posterior distribution parameters (m_x, S_x) from $(m_x^{\setminus i}, S_x^{\setminus i})$ via moment matching. The mean and variance of $\phi_i(\vec{x}_i) p_{\setminus i}(\vec{x})$ can be found by partial differentiation of the log partition function with respect to m and S respectively.
3. The parameters of the factor are updated as:

$$S_i = (S_x^{-1} - (S_x^{\setminus i})^{-1})^{-1} \tag{6.8}$$

$$m_i = S_i (S_x^{-1} m_x - (S_x^{\setminus i})^{-1} m_x^{\setminus i}) \tag{6.9}$$

EP is not popular in computer vision communities because of its computational complexity. For a grid-based Markov Random Field model representing D pixels in an image, EP

requires $O(D^2)$ space required to store the $D \times D$ covariance matrix S_x and a computational complexity limited by the matrix inverse of eq. 6.6 which is $O(D^3)$. EP computations can be made more efficient if the potentials ϕ_i are of reduced rank [5, 42]. The potential ϕ_i is said to be rank K if it can be expressed as $\phi_i(V_i \vec{x})$ where V_i is a $K \times D$ matrix and \vec{x} is a D dimensional vector. It has been shown that it is sufficient to store $V_i S_i V_i'$ and $V_i \vec{m}_i$ instead of the full $D \times D$ covariance matrix S_i . Nonetheless, the covariance matrix of the posterior S remains full-rank with a space complexity of $O(D^2)$. For vision problems, this becomes a limiting factor when dealing with even small images (for e.g. a 256×256 image) by todays computing standards.

This can be alleviated to some extent by exploiting the sparseness of the graph underlying eqn. 6.2 by storing the inverse covariance matrix instead of the covariance matrix. Both the space and computational complexity of each iteration of EP ultimately boils down to the degree of sparsity present in the graph.

6.4 Whitened Expectation Propagation

It has been shown in [5] that even after applying all the above mentioned optimizations, EP remains to be hard for vision applications with a runtime complexity of at least $O(D^{2.5})$. It is preferable for the runtime to scale linearly with the size of the problem. However EP has a desirable property that the complexity does not depend on the clique size, not restricting the model to pairwise interactions.

In [5], the authors propose an EP algorithm that achieves a runtime that is linear in the number of pixels by making use of the knowledge of statistical regularities present in natural imagery. It is a well studied fact that intensity and range images of natural scenes tend to have a power spectrum that obeys a power law $\frac{A}{f^\beta}$ for constants A and β as already detailed in chapter 4. They do it by limiting the forms of covariance structure expressible by S .

$$S^{-1} = \sum \theta_k B_k \quad (6.10)$$

where B_k is a symmetric $D \times D$ matrix and B_k is selected to satisfy three important constraints:

1. it encodes the covariance structure of natural images.
2. the stationary statistics present in natural images and
3. achieve a runtime which is linear in the number of pixels.

For natural images, the covariance matrix will be circulant given its stationary statistics and its eigenvalue decomposition is: $S = FAF'$ where F is a Fourier basis and A is a diagonal matrix whose diagonal is the power spectrum of natural scenes. $\mathcal{W} = FA^{-1/2}F'$ is a whitening matrix for natural scenes which when applied as a filter to an image will remove all correlations that exist between the pixels. In [5], it is shown that a linear runtime can be achieved if the covariance matrix S is constrained to be of the form $\mathcal{W}^{-1}D_S\mathcal{W}^{-1}$ where D_S is a diagonal matrix and allowed to vary. The whitened EP update equations for a potential i are:

$$D_{\setminus i} = (D_S^{-1} - D_i^{-1})^{-1} \quad (6.11)$$

$$\mathcal{W}\mu_{\setminus i} = D_{\setminus i}(D_S^{-1}\mathcal{W}\mu - D_i^{-1}\mathcal{W}\mu_i) \quad (6.12)$$

$$D_S = D_{\setminus i} - D_{\setminus i} \text{diag}[V'_w(I - (V_i S V'_i)(V_w D_{\setminus i} V'_w)^{-1})(V_w D_{\setminus i} V'_w)^{-1} V_w] D_{\setminus i} \quad (6.13)$$

$$\mathcal{W}\mu = \mathcal{W}\mu_{\setminus i} + D_{\setminus i} V'_w (V_w D_{\setminus i} V'_w)^{-1} (V_i \mu - V_i \mu_i) \quad (6.14)$$

$$D_i = (D_S^{-1} - D_{\setminus i}^{-1})^{-1} \quad (6.15)$$

$$\mathcal{W}\mu_i = D_i(D_S^{-1}\mathcal{W}\mu - D_{\setminus i}^{-1}\mathcal{W}\mu_{\setminus i}) \quad (6.16)$$

where $V_w = V_i\mathcal{W}^{-1}$. It is shown in [5] that whitened EP is linear in the number of pixels and clique size.

6.5 Whitened EP for Shape From Shading (SfS)

The goal of SfS is to estimate 3D shape from a single intensity image. It makes assumptions about the scene such as uniform albedo, single point light source of known direction, Lambertian reflectance, no lighting interreflections and cast shadows. But these assumptions are highly unrealistic and traditional SfS techniques generalize poorly to real-world images. In order for them to compete with human capabilities, they have to account for non-Lambertian reflectance, shadow cues, unknown lighting and albedo. This requires MRF models with fully connected potentials which can capture the rich statistical structure present in real-world imagery. Until recently, probabilistic inference on MRF's with large cliques for vision was impractical. Optimized inference algorithms like Sparse EP and Whitened EP [5] will allow us to use MRF models with higher order, non pairwise potentials which will hopefully generalize SfS well to real-world imagery.

Let $i(x, y)$ be the input intensity image and $z(x, y)$ be the corresponding range image. The gradients are defined as $p = \frac{\partial z}{\partial x}$ and $q = \frac{\partial z}{\partial y}$ and $i(x, y) = R(p, q)$ where R is the reflectance function which is traditionally assumed to be Lambertian. The MRF model used for SfS in [5] uses a Gaussian prior on z with zero mean and covariance matrix S equal to the covariance structure of natural images. Following is the equation for the model:

$$P(z) \propto \exp\left(-\frac{1}{2}z'S^{-1}z\right) \prod_{x,y} \phi_R(p_{x,y}, q_{x,y}|i_{x,y}) \quad (6.17)$$

where S is the covariance matrix given by an $\frac{A}{f^\beta}$ power spectrum power law for range

images; where A is the strength of the spatial prior. The potential $\phi_R(p_{x,y}, q_{x,y}|i_{x,y})$ denotes the unnormalized joint likelihood of the gradients p and q given intensity i at pixel location (x, y) .

6.6 Implementation details

A Gaussian prior was used for z with a covariance matrix S whose diagonal is the power spectrum of the range images. In [5], a $1/f^2$ covariance structure was used to be consistent with the covariance statistics of range images of cluttered scenes. But based on the results seen in chapter 4, a $1/f^4$ covariance structure is chosen to model the covariance of single unoccluded objects.

6.6.1 My approach: Learning potentials

In [5], the authors use a Laplace distribution to model ϕ_R to penalize depth maps z that are not consistent with the known pixel intensity:

$$\phi_R(p_{x,y}, q_{x,y}|i(x, y)) = e^{-|R(p_{x,y}, q_{x,y}) - i(x, y)|/b} \quad (6.18)$$

where $R(p, q)$ is the predicted reflectance map and i is the ground truth intensity at pixel (x, y) . For Lambertian reflectance, R is computed as:

$$R(p, q) = \rho \vec{N} \cdot \vec{s} \quad (6.19)$$

where ρ is the albedo of the surface, \vec{N} is the unit normal to the surface given by:

$$\vec{N} = \frac{1}{\sqrt{1 + p^2 + q^2}}(-p, -q, 1)^T \quad (6.20)$$

and \vec{s} is the unit vector pointing towards the light source which is assumed to be known:

$$\vec{s} = \frac{1}{\sqrt{1 + p_s^2 + q_s^2}}(-p_s, -q_s, 1)^T \quad (6.21)$$

and hence the reflectance map is:

$$R(p, q) = \frac{\rho(1 + pp_s + qq_s)}{\sqrt{(1 + p^2 + q^2)}\sqrt{(1 + p_s^2 + q_s^2)}} \quad (6.22)$$

They tested their shape inference algorithm even for non-Lambertian reflectances and showed consistent results. However, there is an assumption of known reflectance properties which is built into the functional form of the potentials. In contrast, my approach is to learn the potentials directly from the data without any prior assumption of surface reflectance. The potentials $\phi_R(p, q|i)$ are learned directly from the SORH dataset of range images using a patch based approach.

Firstly, a dictionary of 25×25 image patches (both intensity and corresponding range) was constructed from the SORH dataset. The dictionary contains image patches which are within the object boundary entirely as well as patches which fall near the occlusion boundaries of the objects. Such image patches contain invalid pixels which correspond to those that are outside the object and valid pixels that fall on/within the object boundaries. The intent is to make sure we have enough samples to account for p, q distributions corresponding to within object shape gradations as well as occlusion boundaries. Potentials are always learned from images taken under the same lighting direction as changes in lighting will result in different intensity variations for the same 3D surface. For each pixel in the intensity image, a 25×25 patch centered at that pixel is taken and matched with the dictionary image patches in the database using k Nearest Neighbor (kNN) search, to get the top $k = 1000$ intensity patch matches along with the corresponding range patches. Using a good similarity metric is crucial to the accuracy of the learned potentials and considerable time was spent testing various metrics for our purpose.

6.6.1.1 Similarity measures

Computing the similarity between two image patches is a common task in computer vision. There is no one universal metric which works well for all scenarios, but needs to be adapted or designed as per the needs of the specific application. Let $X = x_i : i = 1, 2, \dots, n$ and $Y = y_i : i = 1, 2, \dots, n$ represent the intensities of two vectorized image patches containing n pixels. The most popular similarity measures are:

L_1 **norm** or Manhattan distance is the simplest measure of similarity between two image patches. It is the sum of absolute intensity differences between two image patches. Mathematically it is defined as:

$$L_1 = \sum_{i=1}^n |x_i - y_i| \quad (6.23)$$

L_2 **norm** or Euclidean distance is the square root of the sum of squared intensity differences of corresponding pixels. Mathematically it is defined as:

$$L_2 = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (6.24)$$

L_1 and L_2 similarity measures compare raw intensities of corresponding pixel pairs and are highly sensitive to even slight fluctuations in lighting and sensor noise.

Pearson's Correlation Coefficient addresses these issues and is defined as:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (6.25)$$

	Rabbit			Fossil		
	Linear regression	$1/f^3$ EP	$1/f^4$ EP	Linear regression	$1/f^3$ EP	$1/f^4$ EP
1°	0.7	0.99	0.38	0.3	0.4	0.19
2°	2.4	3.85	1.79	0.9	1.38	0.95
3°	5.1	7.55	4.73	1.8	3.43	2.52
4°	8.4	12.35	9.54	3.1	6.58	5
5°	12.2	17.9	14.98	5	10.75	8.3
10°	33.6	43.68	41.38	23	34.94	31.36
15°	51.4	63.4	61.8	43.2	57.57	54.52
20°	65.5	79.3	78.38	61.3	78.3	77
25°	77.4	85.6	86.25	77.9	89.8	89.8

Table 6.1: Comparison of results for linear regression and whitened EP. Each row corresponds to the percentage of surface normals that lie within the given angular separation of the ground-truth surface normal.

where $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ and $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. The equation can also be expressed as follows:

$$r = \frac{1}{n} \sum_{i=1}^n \left(\frac{(x_i - \bar{x})}{\sigma_x} \right) \left(\frac{(y_i - \bar{y})}{\sigma_y} \right) \quad (6.26)$$

where $\sigma_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$ and $\sigma_y = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$ are the standard deviations of X and Y respectively. Pixel intensities are normalized by subtracting the mean and dividing by the standard deviation. This makes the measure invariant to bias in image intensities and contrast changes.

The similarity measures discussed above may be suitable in a typical image matching scenario where all the pixels in a patch are valid intensities. In our case we are faced with two kinds of patches:

- Patches whose pixels are all valid. These correspond to patches that fall entirely within object boundaries.
- Patches which contain both valid and invalid pixels. These correspond to patches that fall near/on the boundary. Invalid pixels are those that are outside the object boundary and patches contain valid and invalid pixels separated by the object boundary.

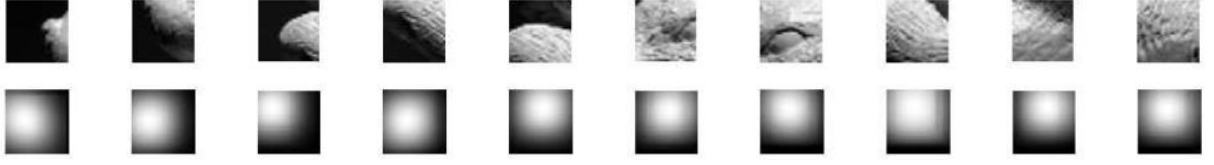


Figure 6.2: Top row: Sample intensity patches. Bottom row: Kernel density estimates for the pixel centered at the corresponding patch in the top row.

Simple pixel-pixel comparison based similarity measures are in fact not sufficient and will result in bad matches. We designed a measure that goes beyond just pixel intensities and quantifies the structure depicted in the image patches. Let $A = a_i : i = 1, 2, \dots, n$ and $B = b_i : i = 1, 2, \dots, n$ represent boolean equivalents of image patches X and Y respectively where:

$$a_i(\text{orb}_i) = \begin{cases} 1 & \text{for valid pixel} \\ 0 & \text{otherwise} \end{cases}$$

The boolean patch describes the image patch in terms of its structure:

- Is it completely inside the object?
- Does it fall near the object boundary? If so, how much of it is outside the boundary and vice versa?

A simple technique is to estimate the amount of overlap of structure between two patches. Specifically, two pixels are overlapping if both are either valid or invalid, $a_i = b_i$ at the same time. We design a new measure o which is the ratio of the number of overlapping pixels to the total number of pixels in the patch.

$$o = \frac{\sum_{i=1}^n a_i \leftrightarrow b_i}{n} \tag{6.27}$$

where the symbol \leftrightarrow denotes logical equality.

We designed a novel similarity measure on the lines of the Structural SIMilarity (SSIM) index [106]. It is a composite of three measures which estimate the overlap of valid pixels between both patches given by eq. 6.27, contrast changes, and correlation respectively. It is defined as:

$$SSIM = [o]^\alpha [c]^\beta [s]^\gamma \quad (6.28)$$

where $SSIM$ is the similarity index computed between two patches, o quantifies the amount of structural overlap and is defined as in eq 6.27, c measures contrast differences and s the correlation which is actually nothing but the Pearson's correlation coefficient redesigned to deal with patches containing invalid pixels. $\alpha > 0$, $\beta > 0$ and $\gamma > 0$ control the relative significance of each of the three terms in the index. Mathematically c and s are defined as follows:

$$c = \frac{2\sigma_x\sigma_y + C_1}{\sigma_x^2 + \sigma_y^2 + C_1} \quad (6.29)$$

$$s = \frac{\sigma_{xy} + C_2}{\sigma_x\sigma_y + C_2} \quad (6.30)$$

where σ_x , σ_y represent the standard deviations of the two patches and σ_{xy} represents the covariance between the patches. The constants C_1 and C_2 are introduced to deal with cases where denominators get close to zero.

Fig. 6.1 shows the matches found by our custom similarity measure on a query patch that falls near the object boundary.

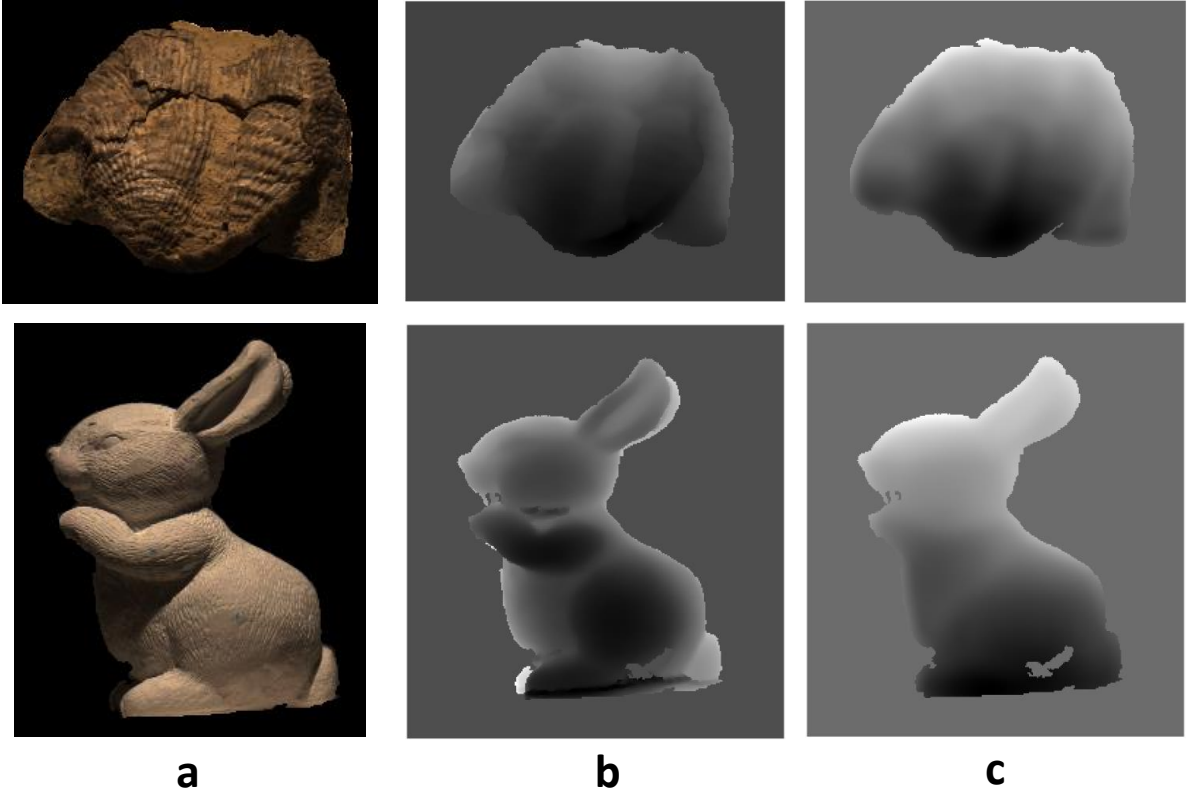


Figure 6.3: a) Color images of test SORH objects, b) ground truth range images, c) predicted range images using $1/f^3$ EP

The similarity values are weighted using a 1D Gaussian vector such that most similar patches are highly weighted than the less similar ones. Gradients p and q are computed at the mid pixels of the range patches which are then used to compute a 2D histogram by accumulating the top 1000 Gaussian weighted similarity values, which denotes the joint distribution of shape and intensity at that pixel. Since the histograms are populated with only a few hundred samples, which come from just a small number of images, they are rough and not very good estimates of the true population joint distribution. We addressed this problem by smoothing the histogram by convolving it with a Gaussian, which is equivalent to kernel density estimation with a Gaussian kernel. The smoothed histogram is then normalized. Potential functions are calculated for all valid pixels in this manner. The kernel

density estimate of the potential is given by:

$$\phi_{KDE}(p, q|i) = \frac{1}{2\pi N \sigma^2} \sum_{i=1}^N s_i e^{\frac{-1}{2\sigma^2}[(p-p_i)^2 + (q-q_i)^2]} \quad (6.31)$$

where s_i is the similarity value of the patch having gradients (p_i, q_i) at its center, N is the number of samples in the histogram which is 1000 and σ is the standard deviation of the Gaussian kernel.

For pixels falling near the object boundaries, the corresponding patches are partially inside/outside the object boundary. The potential functions for such pixels capture the distributions of shape and intensity at boundaries.

Experiments were conducted for both $1/f^3$ and $1/f^4$ covariance structures with prior strengths $A = 300$ and $A = 15000$ respectively which were selected by cross validation and EP was run for 10 iterations. For some potentials, the approximating Gaussian was initially far away from any of the samples in our learned potentials. This can be due to absence of samples in the histogram for the granularity of bins used for p and q . This was resolved by zooming out of the histogram (essentially by increasing the granularity of binning) until the algorithm sees atleast 1 weighted sample. Fig 6.2 shows sample patches and the kernel density estimates for the pixel centered at the corresponding patches. Table. 6.1 shows the percentage of surface normals that lie within the given angular separation of the ground-truth surface normal. The results of whitened EP are compared against shape estimation using linear regression.

6.7 Discussion and open questions

We show that whitened Expectation Propagation for shape inference using a $1/f^3$ covariance structure for single objects with some self occlusion and non parametric density estimation even with a small sample size, shows reasonable results and consistently outper-

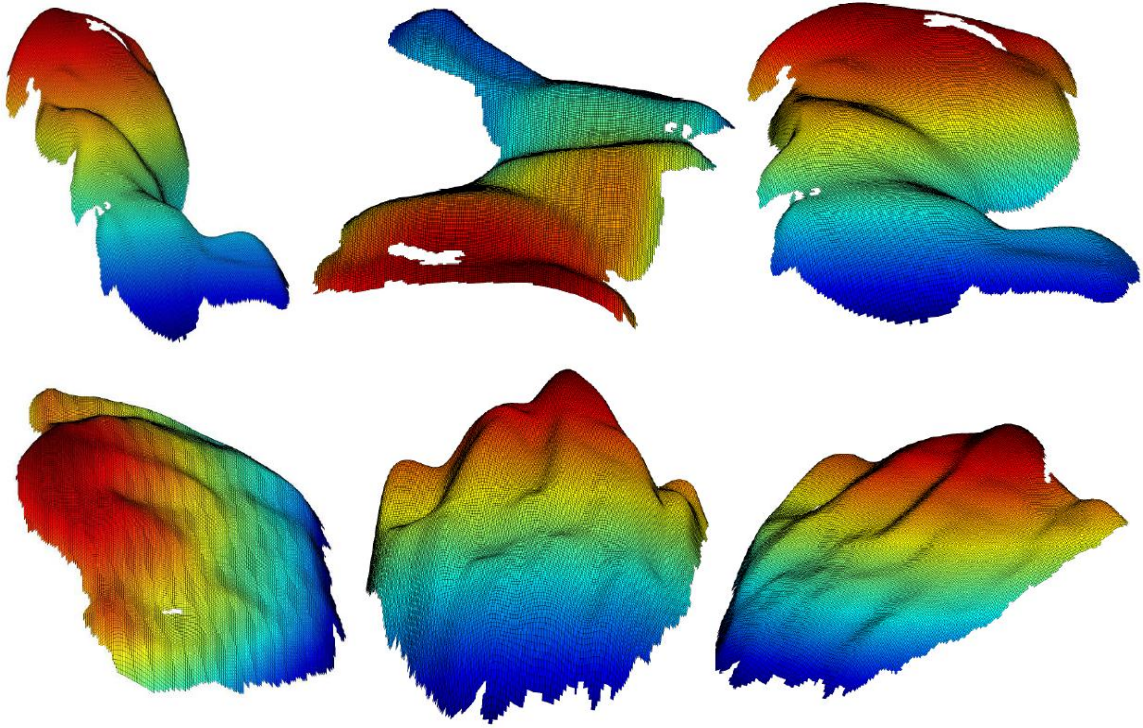


Figure 6.4: 3D mesh plots of predicted shape using EP using a $1/f^3$ covariance structure, for two objects from the SORH database in 3 rotated views.

forms the simple linear regression based shape inference. Non parametric density estimation has an advantage of not having to make assumptions about the form of the data and hence reflects distributions as it appears in the real world. Still, learning accurate potentials from the data requires a considerably large sample size. EP was tested using both $1/f^3$ and $1/f^4$ covariance structures and table 6.1 compares the performance of both. Figs. 6.4 and 6.5 show the 3D mesh plots of predicted shape for EP using the $1/f^3$ and $1/f^4$ covariance structures respectively. The predicted shape for the bunny in fig. 6.5 is flat overall (the range of depth values is less) when compared to the results for EP using $1/f^3$ covariance structure in fig. 6.4. Fig. 6.3a and appendix A, show that the SORH objects are not completely devoid of any occlusion. They exhibit some amount of self occlusion which results in shadowing in the presence of oblique lighting. Hence the use of $1/f^3$ for shape inference for the test objects

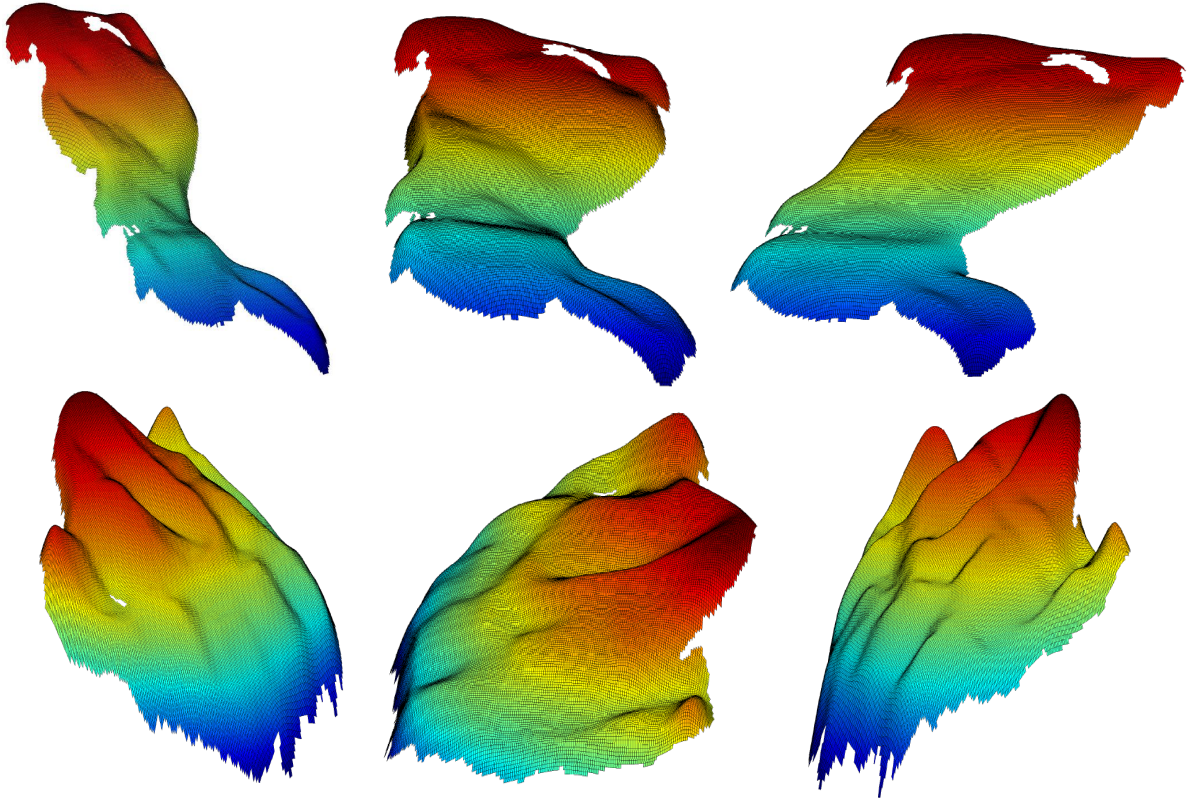


Figure 6.5: 3D mesh plots of predicted shape using EP using a $1/f^4$ covariance structure, for two objects from the SORH database in 3 rotated views.

is justified.

Shape from Shading is a highly ambiguous problem, and any convex shape has a corresponding concave shape that would look identical under the same lighting. The model used didn't explicitly encourage convexity, so the convexity in the results obtained is probably due to the careful handling of the similarity measure using SSIM at the boundaries, where pixels near the edges of surfaces are given surface normals that resemble surface normals seen similarly near surface edges in the training examples. There are still lots of open questions and possible directions for future research. Whitened EP allows us to use any linear filter as the basis for the distributions $\phi(p, q|i)$. We computed p and q as a difference of neighboring range pixel locations. A possible choice would be to compute them at coarser resolutions.

Potentials $\phi(p, q|i)$ are computed at each valid intensity pixel location on the object. We could experiment by taking a block based approach and compute a single potential for the average pixel intensity within that block. Another possible direction would be to try and include additional features in the potential. For example, we can explicitly include occlusion boundary cues. One approach to quantify boundary cues is to compute the distance of the pixel to its nearest boundary as well as the nearest medial (skeletal) point. The skeleton of the object can be computed using a technique called grassfire transforms in image processing. Intuitively, it can be described as "setting fire" at the object borders and the fire would meet at the medial region. The potential $\phi(p, q|i)$ can be replaced with $\phi(p, q|i, d)$ or even $\phi(p, q|i, d, m)$ where d denotes the distance to the nearest boundary and m denotes the distance to the medial axis or skeleton of the object.

References

- [1] Ramachandran and S. Vilayanur, “Perceiving Shape from Shading,” *Scientific American*, vol. 259, pp. 76–83, 1988.
- [2] R. Woodham, “Photometric method for determining surface orientation from multiple images,” vol. 19, pp. 139–144, January 1980.
- [3] B. K. P. Horn, “Shape from shading,” ch. Obtaining shape from shading information, pp. 123–171, Cambridge, MA, USA: MIT Press, 1989.
- [4] B. Potetz and T. S. Lee, “Statistical correlations between two-dimensional images and three-dimensional structures in natural scenes,” *Journal of the Optical Society of America A*, vol. 20, pp. 1292–1303, Jul 2003.
- [5] B. Potetz and M. Hajiabadi, “Whitened expectation propagation: Non-lambertian shape from shading and shadow,” *2013 IEEE Conference on Computer Vision and Pattern Recognition*, vol. 0, pp. 1674–1681, 2013.
- [6] B. A. Olshausen and D. J. Field, “Sparse coding with an overcomplete basis set: a strategy employed by v1?,” *Vision research*, vol. 37, pp. 3311–3325, December 1997.
- [7] D. L. Ruderman and W. Bialek, “Statistics of natural images: scaling in the woods,” *Physical Review Letters*, vol. 73, pp. 814–817, 1994.

- [8] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Journal of Optical Society of America*, vol. 4, no. 12, pp. 2379–2394, 1987.
- [9] A. J. Bell and T. J. Sejnowski, "The "independent components" of natural scenes are edge filters.," *Vision research*, vol. 37, pp. 3327–3338, December 1997.
- [10] Y. Karklin and M. Lewicki, "Emergence of complex cell properties by learning to generalize in natural scenes," *Nature*, vol. 457, pp. 83–86, January 2009.
- [11] Y. Karklin and M. S. Lewicki, "A hierarchical bayesian model for learning non-linear statistical regularities in non-stationary natural signals," *Neural Computation*, vol. 17, no. 2, pp. 397–423, 2005.
- [12] J. Huang and D. Mumford, "Statistics of natural images and models.," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 541–547, 1999.
- [13] J. Huang, A. B. Lee, and D. Mumford, "Statistics of range images.," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1324–1331, 2000.
- [14] A. B. Lee, D. Mumford, and J. Huang, "Occlusion models for natural images: A statistical study of a scale-invariant dead leaves model," *Int. J. Comput. Vision*, vol. 41, no. 1-2, pp. 35–59, 2001.
- [15] Z. Yang and D. Purves, "Image/source statistics of surfaces in natural scenes.," *Network*, vol. 14, pp. 371–390, 2003.
- [16] C. Q. Howe and D. Purves, "Range image statistics can explain the anomalous perception of length," *Proc. Nat. Acad. Sci.*, vol. 99, pp. 13184–13188, 2002.
- [17] A. Torralba and A. Oliva, "Depth estimation from image structure," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, pp. 1226–1238, September 2002.

- [18] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.
- [19] W. T. Freeman and A. Torralba, “Shape recipes: Scene representations that refer to the image,” in *NIPS’02*, pp. 1335–1342, 2002.
- [20] A. Torralba and W. T. Freeman, “Properties and applications of shape recipes,” in *CVPR (2)’03*, pp. 383–390, 2003.
- [21] B. Potetz and T. S. Lee, “Scaling laws in natural scenes and the inference of 3D shape,” in *Advances in Neural Information Processing Systems 18*, pp. 1089–1096, 2006.
- [22] M. J. Brooks, *Shape from shading*. Cambridge, MA, USA: MIT Press, 1989.
- [23] E. Micheli-Tzanakou, X. Shen, and L. Yang, “Shape from shading models,” in *Handbook of Biomedical Image Analysis* (J. S. Suri, D. L. Wilson, and S. Laxminarayan, eds.), Topics in Biomedical Engineering. International Book Series, pp. 257–304, Springer US, 2005.
- [24] R. Zhang, P.-S. Tsai, J. E. Cryer, and M. Shah, “Shape from shading: A survey,” *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, vol. 21, no. 8, pp. 690–706, 1999.
- [25] E. Prados and O. Faugeras, “A generic and provably convergent shape-from-shading method for orthographic and pinhole cameras,” *Int. J. Comput. Vision*, vol. 65, pp. 97–125, November 2005.
- [26] Q. Zheng and R. Chellappa, “Estimation of illuminant direction, albedo, and shape from shading,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 13, pp. 680–702, jul 1991.

- [27] K. M. Lee and C. C. J. Kuo, "Shape from shading with a linear triangular element surface model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, pp. 815–822, August 1993.
- [28] D. Hoiem, A. Efros, and M. Hebert, "Automatic photo pop-up," in *ACM SIGGRAPH*, August 2005.
- [29] D. Hoiem, A. Efros, and M. Hebert, "Closing the loop on scene interpretation," in *Proc. Computer Vision and Pattern Recognition (CVPR)*, June 2008.
- [30] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *NIPS*, 2005.
- [31] A. Saxena, M. Sun, and A. Y. Ng, "Make3d: Learning 3d scene structure from a single still image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 5, pp. 824–840, 2009.
- [32] A. Pentland, "Linear shape from shading," vol. 4, pp. 153–162, March 1990.
- [33] P. sing Tsai and M. Shah, "Shape from shading using linear approximation," *Image and Vision Computing*, vol. 12, pp. 487–498.
- [34] P. Mamassian, D. C. Knill, and D. Kersten, "The perception of cast shadows," *Trends in Cognitive Sciences*, vol. 2, pp. 288–295, Aug. 1998.
- [35] M. S. Langer and H. H. BÄ₄lthoff, "Perception of shape from shading on a cloudy day," *JOURNAL OF THE OPTICAL SOCIETY OF AMERICA A*, vol. 11, no. 11, pp. 467–478, 1999.
- [36] C. W. Tyler, "Diffuse illumination as a default assumption for shape-from-shading in the absence of shadows," in *Human Vision and Electronic Imaging II*, vol. 3016 of *Proceedings of the International Society for Optical Engineering*, (San Jose, California, USA), pp. 346–351, Feb. 1997.

- [37] Y. Yu and J. T. Chang, “Shadow graphs and surface reconstruction,” in *Proceedings of the 7th European Conference on Computer Vision (volume II)*, vol. 2351 of *Lecture Notes in Computer Science*, (Copenhagen, Denmark), pp. 31–45, May 2002.
- [38] M. Daum and G. Dudek, “On 3-d surface reconstruction using shape from shadows,” pp. 461–468, 1998.
- [39] M. S. Langer and S. W. Zucker, “Shape from shading on a cloudy day,” *Journal of the Optical Society of America - Part A: Optics, Image Science, and Vision*, vol. 11, pp. 467–478, Feb. 1994.
- [40] A. J. Stewart and M. S. Langer, “Towards accurate recovery of shape from shading under diffuse lighting,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (San Francisco, California, USA), pp. 411–418, June 1996.
- [41] B. Potetz, “Efficient belief propagation for vision using linear constraint nodes,” in *CVPR 2007: Proceedings of the 2007 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Minneapolis, MN, USA: IEEE Computer Society, 2007.
- [42] T. P. Minka, *A Family of Algorithms for Approximate Bayesian Inference*. PhD thesis, 2001.
- [43] D. Hoiem, A. Efros, and M. Hebert, “Geometric context from a single image,” in *International Conference of Computer Vision (ICCV)*, vol. 1, pp. 654 – 661, IEEE, October 2005.
- [44] J. Sung, C. Ponce, B. Selman, and A. Saxena, “Human activity detection from rgbd images,” in *Association for the Advancement of Artificial Intelligence (AAAI) workshop on Pattern, Activity and Intent Recognition*, 2011.

- [45] J. Sung, B. Selman, and A. Saxena, “Learning sequences of controllers for complex manipulation tasks,” in *International Conference on Machine Learning (ICML) workshop on Prediction with Sequential Models*, 2013.
- [46] Y. Ch  n  ,   tienne Belin, D. Rousseau, and F. Chapeau-Blondeau, “Multiscale analysis of depth images from natural scenes: Scaling in the depth of the woods,” *Chaos, Solitons & Fractals*, vol. 54, pp. 135–149, September 2013.
- [47] S. Winkelbach, S. Molkenstruck, and F. M. Wahl, “Low-cost laser range scanner and fast surface registration approach,” in *DAGM-Symposium*, pp. 718–728, 2006.
- [48] “<http://www.david-laserscanner.com/>.”
- [49] P. E. Debevec and J. Malik, “Recovering high dynamic range radiance maps from photographs,” in *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, SIGGRAPH ’97, pp. 369–378, 1997.
- [50] N. G. Aldrin, T. Zickler, and D. Kriegman, “Photometric stereo with non-parametric and spatially-varying reflectance,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (Anchorage, AK), 2008.
- [51] A. Hertzmann and S. M. Seitz, “Example-based photometric stereo: Shape reconstruction with general, varying brdfs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1254–1264.
- [52] A. Hertzmann and S. M. Seitz, “Shape and materials by example: A photometric stereo approach,” in *IN PROC. IEEE CONF. COMPUTER VISION AND PATTERN RECOGNITION*, pp. 533–540, 2003.
- [53] D. B. Goldman, B. Curless, A. Hertzmann, and S. M. Seitz, “Shape and spatially-varying brdfs from photometric stereo,” in *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1 - Volume 01*, pp. 341–348, 2005.

- [54] “vision.middlebury.edu/stereo/.”
- [55] D. Scharstein and R. Szeliski, “High-accuracy stereo depth maps using structured light,” in *Proceedings of the 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR’03, 2003.
- [56] D. Scharstein and R. Szeliski, “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms,” *Int. J. Comput. Vision*, vol. 47, no. 1-3.
- [57] H. Du, D. B. Goldman, and S. M. Seitz, “Binocular photometric stereo.,” in *BMVC*, pp. 1–11, 2011.
- [58] A. Geiger, P. Lenz, and R. Urtasun, “Are we ready for autonomous driving? the kitti vision benchmark suite,” in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [59] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, “Vision meets robotics: The kitti dataset,” *International Journal of Robotics Research (IJRR)*, 2013.
- [60] S. Meister, B. Jähne, and D. Kondermann, “Outdoor stereo camera system for the generation of real-world benchmark data sets,” *Optical Engineering*, vol. 51, no. 02, p. 021107, 2012.
- [61] H. Chapman, V. Gaffney, and H. L. Moulden, “The Eton Myers collection virtual museum.,” *IJHAC*, vol. 4, no. 1-2, pp. 81–93, 2010.
- [62] A. van der Schaaf and J. van Hateren, “Modelling the power spectra of natural images: Statistics and information,” *Vision Research*, vol. 36, no. 17, pp. 2759 – 2770, 1996.
- [63] R. W. Buccigrossi and E. P. Simoncelli, “Image compression via joint statistical characterization in the wavelet domain,” *IEEE Transactions on Image Processing*, pp. 1688–1701, 1999.

- [64] M. Elad and M. Aharon, “Image denoising via sparse and redundant representations over learned dictionaries,” *Image Processing, IEEE Transactions on*, vol. 15, pp. 3736–3745, November 2006.
- [65] S. Roth and M. J. Black, “Fields of experts: A framework for learning image priors,” in *CVPR*, pp. 860–867, 2005.
- [66] V. Jain and S. Seung, “Natural image denoising with convolutional networks,” in *Advances in Neural Information Processing Systems 21* (D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, eds.), pp. 769–776, 2009.
- [67] M. Heiler and C. Schnörr, “Natural image statistics for natural image segmentation,” in *ICCV ’03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, (Washington, DC, USA), p. 1259, IEEE Computer Society, 2003.
- [68] S. S. Ekhande, S. P. Sonavane, and P. J. Kulkarni, “Article: Universal steganalysis using feature selection strategy for higher order image statistics,” *International Journal of Computer Applications*, vol. 1, pp. 53–56, February 2010.
- [69] T. Pouli, D. Cunningham, and E. Reinhard, “Image statistics and their applications in computer graphics,” in *Eurographics State-of-the-Art-Reports*, 2010.
- [70] P. T. O. Woodford, I. Reid and A. Fitzgibbon, “Fields of experts for image-based rendering,” in *Proceedings British Machine Vision Conference*, 2006.
- [71] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. The MIT Press, December 2001.
- [72] H. Lee, C. Ekanadham, and A. Y. Ng, “Sparse deep belief net model for visual area V2,” in *Advances in Neural Information Processing Systems 20*, pp. 873–880, 2008.
- [73] G. E. Hinton, “What kind of graphical model is the brain?,” in *IJCAI*, pp. 1765–, 2005.

- [74] D. W. Dong and J. J. Atick, “Statistics of natural time-varying images,” 1995.
- [75] R. Grosse, M. K. Johnson, E. H. Adelson, and W. T. Freeman, “Ground-truth dataset and baseline evaluations for intrinsic image algorithms,” in *International Conference on Computer Vision*, pp. 2335–2342, 2009.
- [76] R. O. Dror, T. K. Leung, E. H. Adelson, and A. S. Willsky, “Statistics of real-world illumination,” in *CVPR (2)*, pp. 164–171, 2001.
- [77] D. R. Martin, C. Fowlkes, D. Tal, and J. Malik, “A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics,” Tech. Rep. UCB/CSD-01-1133, EECS Department, University of California, Berkeley, Jan 2001.
- [78] J. M. Samonds, B. Potetz, and T. S. Lee, “Relative luminance and binocular disparity preferences are correlated in macaque v1, matching natural scene statistics,” *Proc Nat Acad Sci USA (PNAS)*, vol. 109, pp. 6313–6318, April 2012.
- [79] D. L. Ruderman, “Origins of scaling in natural images,” *Vision Research*, vol. 37, pp. 3385–3398, 1997.
- [80] G. J. Burton and I. R. Moorhead, “Color and spatial structure in natural scenes,” *Appl. Opt.*, vol. 26, pp. 157–170, Jan 1987.
- [81] D. J. Tolhurst, Y. Tadmor, and T. Chao, “Amplitude spectra of natural images,” *Ophthalmic and Physiological Optics*, vol. 12, no. 2, pp. 229–232, 1992.
- [82] A. Torralba and A. Oliva, “Statistics of natural image categories,” *Network: Computation in Neural Systems*, pp. 391–412, Aug. 2003.
- [83] P. Prusinkiewicz and J. Hanan, *Lindenmayer systems, fractals, and plants*. Lecture notes in biomathematics, Springer-Verlag, 1989.

- [84] A. Lee, D. Mumford, and J. Huang, “Occlusion Models for Natural Images: A Statistical Study of a Scale-Invariant Dead Leaves Model,” *International Journal of Computer Vision*, vol. 41, no. 1, pp. 35–59, 2001.
- [85] D. Stoyan, W. Kendall, and J. Mecke, *Stochastic geometry and its applications*. Wiley series in probability and statistics, 1995.
- [86] G. Matheron, “Modèle séquentiel de partition aléatoire,” *Technical report, CMM*, 1968.
- [87] R. M. Balboa, C. W. Tyler, and N. M. Grzywacz, “Occlusions contribute to scaling in natural images,” *Vision Research*, vol. 41, no. 7, pp. 955–64, 2001. PMID: 11248280.
- [88] N. M. Grzywacz, R. M. Balboa, and C. W. Tyler, “Letter to the editor: Response,” *Vision Research*, vol. 42, pp. 2803–2805, 2002.
- [89] L. Alvarez, Y. Gousseau, and J.-M. Morel, *The Size of Objects in Natural and Artificial Images*, vol. 111, pp. 167–242. Elsevier, 1999.
- [90] M. S. Langer, “Large-scale failures of f - α scaling in natural image spectra,” *J. Opt. Soc. Am. A*, vol. 17, pp. 28–33, Jan 2000.
- [91] Z. Chi, *Probability Models for Complex Systems*. PhD thesis, Brown University, May 1998.
- [92] E. Reinhard, P. Shirley, M. Ashikhmin, and T. Troscianko, “Second order image statistics in computer graphics,” in *Proceedings of the 1st Symposium on Applied perception in graphics and visualization*, APGV ’04, (New York, NY, USA), pp. 99–106, ACM, 2004.
- [93] C. M. Ramsay, “The distribution of sums of certain i.i.d. pareto variates,” *Communications in Statistics - Theory and Methods*, vol. 35, no. 3, pp. 395–405, 2006.

- [94] C. Wilke, S. Altmeyer, and T. Martinetz, “Large-scale evolution and extinction in a hierarchically structured environment,” in *Proceedings of the Sixth International Conference on Artificial Life*, pp. 266–272, MIT Press, 1998.
- [95] B. Pilgram, D. T. Kaplan, and D. T. Kaplan, “A comparison of estimators for 1/f noise,” 1997.
- [96] B. D. Rigling, “Maximum likelihood estimation of band-limited power law spectrums,” *IEEE Signal Process. Lett.*, vol. 19, no. 5, pp. 307–310, 2012.
- [97] R. J. PIKE and W. J. ROZEMA, “Spectral analysis of landforms*,” *Annals of the Association of American Geographers*, vol. 65, no. 4, pp. 499–516, 1975.
- [98] W. H. Hsiao and R. P. Millane, “Effects of occlusion, edges, and scaling on the power spectra of natural images,” *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, vol. 22, no. 9, pp. 1789–97, 2005. PMID: 16211805.
- [99] B. Potetz and M. Hajiabadi, “Whitened expectation propagation: Non-lambertian shape from shading and shadow,” in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pp. 1674–1681, IEEE, 2013.
- [100] P. Kohli, P. H. Torr, *et al.*, “Robust higher order potentials for enforcing label consistency,” *International Journal of Computer Vision*, vol. 82, no. 3, pp. 302–324, 2009.
- [101] O. Woodford, P. Torr, I. Reid, and A. Fitzgibbon, “Global stereo reconstruction under second-order smoothness priors,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 12, pp. 2115–2128, 2009.
- [102] P. F. Felzenszwalb and D. P. Huttenlocher, “Efficient belief propagation for early vision,” *Int. J. Comput. Vision*, vol. 70, pp. 41–54, Oct. 2006.

- [103] J. Sun, N.-N. Zheng, and H.-Y. Shum, "Stereo matching using belief propagation," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 7, pp. 787–800, 2003.
- [104] X. Lan, S. Roth, D. Huttenlocher, and M. J. Black, "Efficient belief propagation with learned higher-order markov random fields," in *Computer Vision–ECCV 2006*, pp. 269–282, 2006.
- [105] T. P. Minka, "Expectation propagation for approximate bayesian inference," in *Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence*, UAI '01, pp. 362–369, 2001.
- [106] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Perceptual image quality assessment: From error visibility to structural similarity," *IEEE Trans Image Processing*, vol. 13, pp. 600–612, April 2004.

Appendix A

SORH Imagery



